

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

REMARKS

Claims 21-40 are pending in the application. Applicants reserve the right to prosecute non-elected subject matter in subsequent divisional applications.

Comments Regarding Restriction Requirement

Applicants affirm the election with traverse of Group II, which corresponds to newly added claims 23-31 drawn to a polynucleotide, vector, host cell, and method for producing a polypeptide. Newly added claims 23-31 replace original claims 3-6 and 9-14, and are drawn to substantially the same invention, but are of a different scope.

Applicants respectfully submit that there is minimal additional burden on the Examiner to examine newly added claims 39 and 40, which are drawn to microarrays using the elected polynucleotides.

Applicants request that the Examiner withdraw the Restriction Requirement at least with respect to claims 21, 22, 35, and 36 of Group I, and examine those claims together with the elected polynucleotide claims of Group II.

The rules under MPEP section 1893.03(d) require the Examiner to apply the Unity of Invention standard PCT Rule 13.2 instead of U.S. restriction/election of species practice in national stage applications, such as the instant application filed under 35 U.S.C. 371. Applicants believe unity of invention exists for claims drawn to the polypeptide sequence of SEQ ID NO:1 (*i.e.*, claims 21, 22, 35, and 36) and claims drawn to the elected polynucleotide sequence of SEQ ID NO:2 which encodes SEQ ID NO:1 (*i.e.*, claims 23-31) based on the rules concerning unity of invention under the Patent Cooperation Treaty. The Administrative Instructions Under The Patent Cooperation Treaty, Annex B, Unity of Invention, Part 2, "Examples Concerning Unity of Invention" provide the following guidelines with regard to unity of invention between a protein and the polynucleotide that encodes it:

Example 17

Claim 1: Protein X.

Claim 2: DNA sequence encoding protein X.

Expression of the DNA sequence in a host results in the production of a protein which is determined by the DNA sequence. The protein and the DNA sequence exhibit corresponding special technical features. Unity between claims 1 and 2 is accepted.

As currently pending, the claims of Group II drawn to polynucleotides and the claims of Group I drawn to polypeptides do not encompass prior art, and the “objection of lack of unity” based on the reference of Kinkema et al. (Accession Q39157) no longer applies. Therefore, Applicants request that the Examiner withdraw the Restriction Requirement, at least with respect to claims 21, 22, 35, and 36 of Group I, and examine those claims together with the elected polynucleotide claims of Group II.

Rejoinder of method claims upon allowance of product claims under U.S. practice

The Examiner is reminded that claims 32-34 and 38, drawn to methods of using the elected polynucleotides of Group II should be rejoined per the Commissioner’s Notice in the Official Gazette of March 26, 1996, entitled “Guidance on Treatment of Product and Process Claims in light of *In re Ochiai*, *In re Brouwer* and 35 U.S.C. § 103(b)” which sets forth the rules, upon allowance of product claims, for rejoinder of process claims covering the same scope of products. Applicants request that claims 32-34 and 38 be rejoined and examined upon allowance of the claims drawn to the polynucleotides of Group II.

Objections to the claims

Original claims 3-6 were objected to because of their dependence from original claim 1. New claims 23-26 similarly depend from claim 21, drawn to nonelected polypeptides. As mentioned above, Applicants believe that the claims drawn to the polypeptides of the invention, according to the unity of invention standard, should be examined with the elected claims drawn to the polynucleotides currently under examination. Applicants request reconsideration and believe amending these claims at this time would be premature.

Original claims 4 and 10 were objected to as being in improper dependent form. These claims have now been replaced by new claims 23 and 30, which are believed to be in proper form. Withdrawal of the objections is therefore respectfully requested.

Utility Rejections under 35 U.S.C. §101 and §112, First Paragraph

Original claims 3-6 and 9-14, now replaced by new claims 23-31, have been rejected under 35 U.S.C. §101 and §112, first paragraph, because the claimed invention allegedly “is not supported by either a credible asserted utility or a well-established utility” (Office Action, page 3). These rejections are traversed.

The rejection of claims 23-31 is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well known to one of ordinary skill in the art.

The invention at issue is a polynucleotide sequence corresponding to a gene that is expressed in hematopoietic/immune system, gastrointestinal, musculoskeletal, and reproductive tissues, and in tissues associated with cancer (Specification at page 18, lines 12-17). In particular, similarities between SEQ ID NO:1 and *C. elegans* myosin (g1279777) and *H. annuus* unconventional myosin (g2444174), including the presence of myosin head domain, myosin heavy chain, and light chain binding site signatures, are described in the specification, for example, at page 17, line 30 through page 18, line 9. The specification points out the roles of myosin in muscle contraction, intracellular movement, phagocytosis, and cytokinesis, and describes various diseases associated with myosin dysfunction, including muscle disorders, cardiovascular disease, deafness, and cancer (Specification at pages 1-2). As such, the claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires knowledge of how the polypeptide coded for by the polynucleotide actually functions.

Applicants submit with this paper the Declaration of Dr. Tod Bedilion¹ describing some of the practical uses of the claimed invention in gene and protein expression monitoring applications. The Bedilion Declaration demonstrates that the positions and arguments made by the Patent Examiner with respect to the utility of the claimed polynucleotide are without merit.

¹The Bedilion Declaration is submitted herewith in unexecuted form. The executed Declaration will be submitted to the Patent office as soon it is available.

The Bedilion Declaration describes, in particular, how the claimed expressed polynucleotide can be used in gene expression monitoring applications that were well-known at the time the patent application was filed, and how those applications are useful in developing drugs and monitoring their activity. Dr. Bedilion states that the claimed invention is a useful tool when employed as a highly specific probe in a cDNA microarray:

Persons skilled in the art would appreciate that cDNA microarrays that contained the SEQ ID NO:1-encoding polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer for such purposes as evaluating their efficacy and toxicity.

The Patent Examiner does not dispute that the claimed polynucleotide can be used as a probe in cDNA microarrays and used in gene expression monitoring applications. Instead, the Patent Examiner contends that the claimed polynucleotide cannot be useful without precise knowledge of its biological function. But the law never has required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotide in the absence of any knowledge as to the precise function of the protein encoded by it. The uses of the claimed polynucleotide in gene expression monitoring applications are in fact independent of its precise function.

I. The Applicable Legal Standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

An invention is “useful” under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) (“to violate Section 101 the claimed device must be totally incapable of achieving a useful result”); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention “is incapable of serving any beneficial end”).

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: “[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility.” *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See *Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a “nebulous expression” such as “biological activity” or “biological properties” that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be “substantial.” *Brenner*, 383 U.S. at 534. A “substantial” utility is a practical, “real-world” utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a “well-established” utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examination Procedure at § 706.03(a). Only if there is no “well-established” utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*,

51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a “substantial likelihood” of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

II. Use of the claimed polynucleotide for diagnosis of conditions or diseases characterized by expression of MHCH, for toxicology testing, and for drug discovery are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are “well-established” uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application’s specification. These uses are explained, in detail, in the Bedilion Declaration accompanying this response. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

A. The use of MHCH for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer “specific benefits” to the public

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotide.

In his Declaration, Dr. Bedilion explains the many reasons why a person skilled in the art reading the Tang ‘248 application on November 5, 1998 would have understood that application to disclose the claimed polynucleotide to be useful for a number of gene expression monitoring

applications, *e.g.*, as a highly specific probe for the expression of that specific polynucleotide in connection with the development of drugs and the monitoring of the activity of such drugs. (Bedilion Declaration at, *e.g.*, ¶¶ 10-15). Much, but not all, of Dr. Bedilion's explanation concerns the use of the claimed polynucleotide in cDNA microarrays of the type first developed at Stanford University for evaluating the efficacy and toxicity of drugs, as well as for other applications. (Bedilion Declaration, ¶¶ 12 and 15).²

In connection with his explanations, Dr. Bedilion states that the "Tang '248 specification would have led a person skilled in the art on November 5, 1998 who was using gene expression monitoring in connection with working on developing new drugs for the treatment of heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer [a] to conclude that a cDNA microarray that contained the SEQ ID NO:1-encoding polynucleotides would be a highly useful tool, and [b] to request specifically that any cDNA microarray that was being used for such purposes contain the SEQ ID NO:1-encoding polynucleotides" (Bedilion Declaration, ¶ 15). For example, as explained by Dr. Bedilion, "[p]ersons skilled in the art would [have appreciated on November 5, 1998] that a cDNA microarray that contained the SEQ ID NO:1-encoding polynucleotides would be a more useful tool than a cDNA microarray that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer for such purposes as evaluating their efficacy and toxicity." *Id.*

In support of those statements, Dr. Bedilion provided detailed explanations of how cDNA technology can be used to conduct gene expression monitoring evaluations, with extensive citations to pre-November 5, 1998 publications showing the state of the art on November 5, 1998. (Bedilion Declaration, ¶¶ 10-14). While Dr. Bedilion's explanations in paragraph 15 of his Declaration include almost three pages of text and six subparts (a)-(f), he specifically states that his explanations are not "all-inclusive." *Id.* For example, with respect to toxicity evaluations, Dr. Bedilion had earlier explained

²Dr. Bedilion also explained, for example, why persons skilled in the art would also appreciate, based on the Tang '248 specification, that the claimed polynucleotide would be useful in connection with developing new drugs using technology, such as Northern analysis, that predated by many years the development of the cDNA technology (Bedilion Declaration, ¶ 16).

how persons skilled in the art who were working on drug development on November 5, 1998 (and for several years prior to November 5, 1998) “without any doubt” appreciated that the toxicity (or lack of toxicity) of any proposed drug was “one of the most important criteria to be evaluated in connection with the development of the drug” and how the teachings of the Tang ‘248 application clearly include using differential gene expression analyses in toxicity studies (Bedilion Declaration, ¶ 10).

Thus, the Bedilion Declaration establishes that persons skilled in the art reading the Tang ‘248 application at the time it was filed “would have wanted their cDNA microarray to have a [SEQ ID NO:1-encoding polynucleotide probe] because a microarray that contained such a probe (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies using cDNA microarrays that persons skilled in the art have been doing since well prior to November 5, 1998” (Bedilion Declaration, ¶ 15, item (f)). This, by itself, provides more than sufficient reason to compel the conclusion that the Tang ‘248 application disclosed to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the claimed polynucleotide.

Nowhere does the Patent Examiner address the fact that, as described on pp. 31-32 of the Tang ‘248 application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays – probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon v. Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 (“Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)” (emphasis added)).

Though Applicants need not so prove to demonstrate utility, there can be no reasonable dispute that persons of ordinary skill in the art have numerous uses for information about relative gene expression including, for example, understanding the effects of a potential drug for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer. Because the patent application states explicitly that the claimed polynucleotide is known to be expressed in hematopoietic/immune system, gastrointestinal, musculoskeletal, and reproductive tissues, and in tissues associated with cancer (Specification at page 18, lines 12-17), and expresses a protein that is a member of the myosin family known to be associated with diseases such as heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer, there can be no reasonable dispute that a person of ordinary skill in the art could put the claimed invention to such use. In other words, the person of ordinary skill in the art can derive more information about a potential heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer drug candidate or potential toxin with the claimed invention than without it (see Bedilion Declaration at, e.g., ¶ 15, subparts (e)-(f)).

The Bedilion Declaration shows that a number of pre-November 5, 1998 publications confirm and further establish the utility of cDNA microarrays in a wide range of drug development gene expression monitoring applications at the time the Tang '248 application was filed (Bedilion Declaration ¶¶ 10-14; Bedilion Exhibits A-G). Indeed, Brown and Shalon U.S. Patent No. 5,807,522 (the Brown '522 patent, Bedilion Exhibit D), which issued from a patent application filed in June 1995 and was effectively published on December 29, 1995 as a result of the publication of a PCT counterpart application, shows that the Patent Office recognizes the patentable utility of the cDNA technology developed in the early to mid-1990s. As explained by Dr. Bedilion, among other things (Bedilion Declaration, ¶ 12):

The Brown '522 patent further teaches that the “[m]icroarrays of immobilized nucleic acid sequences prepared in accordance with the invention” can be used in “numerous” genetic applications, including “monitoring of gene expression” applications (see Bedilion Tab D at col. 14, lines 36-42). The Brown '522 patent teaches (a) monitoring gene expression (i) in different tissue types, (ii) in different disease states, and (iii) in response to different drugs, and (b) that arrays disclosed therein may be used in toxicology studies (see Bedilion Tab D at col. 15, lines 13-18 and 52-58 and col. 18, lines 25-30).

Literature reviews published shortly after the filing of the Tang '248 application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

* * *

Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

* * *

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis added)

Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, 29 Xenobiotica No. 7, 655 (1999).

In another pre-November 5, 1998 article, Lashkari et al. state explicitly that sequences that are merely “predicted” to be expressed (predicted Open Reading Frames, or ORFs) – the claimed invention in fact is known to be expressed – have numerous uses:

Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons– they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay.

Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, 94 Proc. Nat. Acad. Sci. 8945 (Aug. 1997) (emphasis added).

B. The use of nucleic acids coding for proteins expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now “well-established”

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, as described by Bedilion in his Declaration.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., *supra*:

Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs.

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and Toxicology: The Advent of Toxicogenomics, 24 Molecular Carcinogenesis 153 (1999); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, 112-13 Toxicology Letters 467 (2000).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. “Arrays are at their most powerful when they contain the entire genome of the species they are being used to study.” John C. Rockett and David J. Dix, Application of DNA Arrays to Toxicology, 107 *Environ. Health Perspec.* 681, No. 8 (1999). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding, indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

In fact, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Recent developments provide evidence that the benefits of this information are already beginning to manifest themselves. Examples include the following:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.
- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte’s genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.
- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the “well-established” utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner’s rejections should be overturned regardless of their merit.

C. The Uncontested Fact That the Claimed Polynucleotide Encodes for a Protein in the Myosin Family Also Demonstrates Utility

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, it is undisputed that the claimed polynucleotide encodes for a protein having the sequence shown as SEQ ID NO:1 in the patent application and referred to as MHCH in that application. Appellants have demonstrated that MHCH is a member of the myosin family, and that the myosin family of proteins includes motor proteins that are involved in muscle contraction, intracellular movement, phagocytosis, and cytokinesis.

The Patent Examiner does not dispute any of the facts set forth in the previous paragraph. Neither does the Patent Examiner dispute that, if a polynucleotide encodes for a protein that has a substantial, specific and credible utility, then it follows that the polynucleotide also has a substantial, specific and credible utility.

The Examiner must accept the applicant’s demonstration that the polypeptide encoded by the claimed invention is a member of the myosin family and that utility is proven by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

Nor has the Examiner provided any evidence that any member of the myosin family, let alone a substantial number of those members, is not useful. In such circumstances, the only reasonable inference is that the polypeptide encoded by the claimed invention must be useful, like the other members of the myosin family.

D. Objective evidence corroborates the utilities of the claimed invention

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a “real-world” utility exists. Indeed, “real-world” evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Applicants’ assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte’s customers and the scientific community have acknowledged that Incyte’s databases have proven to be valuable in, for example, the identification and development of drug candidates. As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte’s discovery of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

III. The Patent Examiner’s Rejections Are Without Merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established utilities for the claimed polynucleotide are not “specific or substantial” utilities. (Office Action at p. 4). The Examiner is incorrect both as a matter of law and as a matter of fact.

A. The Precise Biological Role Or Function Of An Expressed Polynucleotide Is Not Required To Demonstrate Utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Bedilion Declaration (at, e.g., ¶¶ 10 and 15, Bedilion), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged so much (66 F.R. at 1095):

[T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

B. Membership in a Class of Useful Products Can Be Proof of Utility

Despite the uncontradicted evidence that the claimed polynucleotide encodes a polypeptide in the myosin family, the Examiner refused to impute the utility of the members of the myosin family to MHCH. In the Office Action, the Patent Examiner takes the position that, unless Applicants can identify which particular biological function within the class of myosins is possessed by MHCH, utility cannot be imputed. To demonstrate utility by membership in the class of myosins, the Examiner would require that all myosins possess a “common” utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. *See Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a “general” class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses MHCH as if the general class in which it is included is not the myosin family, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these “general classes” may contain a substantial number of useless members, the myosin family does not. The myosin family is sufficiently specific to rule out any reasonable possibility that MHCH would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the myosin class of proteins has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a “substantial likelihood” that the MHCH encoded by the claimed polynucleotide is useful. It follows that the claimed polynucleotide also is useful.

It is undisputed that known members of the myosin family are motor proteins involved in muscle contraction, intracellular movement, phagocytosis, and cytokinesis. A person of ordinary skill in the art need not know any more about how the claimed invention functions to use it, and the Examiner presents no evidence to the contrary. The Examiner then goes on to assume that the only use for MHCH absent knowledge as to how the myosin actually works is further study of MHCH itself.

Not so. As demonstrated by Applicants, knowledge that MHCH is a myosin is more than sufficient to make it useful for the diagnosis and treatment of heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer. Indeed, MHCH has been shown to be expressed in hematopoietic/immune system, gastrointestinal, musculoskeletal, and reproductive tissues, and in tissues associated with cancer (Specification at page 18, lines 12-17). The Examiner must accept these facts to be true unless the Examiner can provide evidence or sound scientific reasoning to the contrary. But the Examiner has not done so.

C. Because the uses of polynucleotides encoding MHCH in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the invention itself, the claimed invention has substantial utility.

The PTO rejected the claims at issue on the ground that the use of an invention as a tool for research is not a “substantial” use. Because the PTO’s rejection assumes a substantial overstatement of the law, and is incorrect in fact, it must be overturned.

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office has recognized that just because an invention is used in a research setting does not mean that it lacks utility (MPEP § 2107):

Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact “useful” in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm.

The Patent Office’s actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases. These are acknowledged by the PTO’s Training Materials themselves to be useful, as well as DNA sequences used, for example, as markers.

Only a limited subset of research uses are not “substantial” utilities: those in which the only known use for the claimed invention is to be an **object** of further study, thus merely inviting further research. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the only known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945 (“What Applicants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines.”). Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other beneficial use in research.

As used in toxicology testing, drug discovery, and disease diagnosis, the claimed invention has a beneficial use in research other than studying the claimed invention or its protein products. It is a tool, rather than an object, of research. The data generated in gene expression monitoring using the claimed invention as a tool is **not** used merely to study the claimed polynucleotide itself, but rather to study properties of tissues, cells, and potential drug candidates and toxins. Without the claimed invention, the

information regarding the properties of tissues, cells, drug candidates and toxins is less complete. (Bedilion Declaration at ¶ 15.)

The claimed invention has numerous additional uses as a research tool, each of which alone is a “substantial utility.” These include uses such as diagnostic assays (e.g., pages 36-39), chromosomal markers (e.g., pages 39-40), and ligand screening assays (e.g., page 40).

IV. By Requiring the Patent Applicant to Assert a Particular or Unique Utility, the Patent Examination Utility Guidelines and Training Materials Applied by the Patent Examiner Misstate the Law

There is an additional, independent reason to overturn the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: “specific” utilities which meet the statutory requirements, and “general” utilities which do not. The Training Materials define a “specific utility” as follows:

A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as “gene probe” or “chromosome marker” would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between “specific” and “general” utilities by assessing whether the asserted utility is sufficiently “particular,” *i.e.*, unique (Training Materials at p.52) as compared to the “broad class of invention.” (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) (“With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.”)).

Such “unique” or “particular” utilities never have been required by the law. To meet the utility requirement, the invention need only be “practically useful,” *Natta*, 480 F.2d 1 at 1397, and confer a “specific benefit” on the public. *Brenner*, 383 U.S. at 534. Thus, incredible “throwaway” utilities, such as trying to “patent a transgenic mouse by saying it makes great snake food,” do not meet this standard. Karen Hall, Genomic Warfare, *The American Lawyer* 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where “specific utility” is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be “definite,” not particular. *Montedison*, 664 F.2d at 375. Applicant is not aware of any court that has rejected an assertion of utility on the grounds that it is not “particular” or “unique” to the specific invention. Where courts have found utility to be too “general,” it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had “useful biological activity” was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a “particular” type of cancer was determined to satisfy the specificity requirement). “Particularity” is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § II.B.2 (*Montedison*, 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of “general” utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied,

they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. *See supra* § II.B. Thus the Training Materials cannot be applied consistently with the law.

V. To the Extent the Rejection of the Patented Invention under 35 U.S.C. § 112, First Paragraph, Is Based on the Improper Rejection for Lack of Utility under 35 U.S.C. § 101, it Must Be Reversed.

The rejection set forth in the Office Action is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under § 112, first paragraph, is based on the improper allegation of lack of patentable utility under § 101, it fails for the same reasons.

Enablement rejections under 35 U.S.C. § 112, first paragraph

Original claims 3, 4, 9, and 10, now replaced by new claims 23-25, 30, and 31, are rejected for allegedly failing to meet the requirements of 35 U.S.C. § 112, first paragraph, on the grounds that the Specification does not provide an enabling disclosure commensurate in scope with the claims (Office Action pages 4-5). In particular, the Examiner asserts that “searching for the specific nucleotides to change (deletion, insertion, substitution, or combinations thereof) in a polynucleotide to make any polynucleotide of any nucleotide sequence having 70% identity to any polynucleotide encoding SEQ ID NO:1 or any fragment thereof or any polynucleotide having 70% identity to SEQ ID NO:2 or any fragment thereof is well outside the realm of routine experimentation and predictability in the art...” (Office Action, page 5). The Applicants traverse the rejection for at least the following reasons.

As set forth in *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971):

The first paragraph of § 112 requires nothing more than objective enablement. How such a teaching is set forth, either by the use of illustrative examples or by broad terminology, is of no importance.

As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first

paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.

Applicants submit that the disclosure amply enables the claimed invention. First, Applicants respectfully point out that the claims of the instant application are drawn to **naturally-occurring** variants. Thus it is not necessary to screen every conceivable variant which might be made using recombinant methods, as all that is claimed are those variant sequences which are found in nature. Given the sequences of SEQ ID NO:1 and SEQ ID NO:2, one of ordinary skill in the art could readily identify a polynucleotide encoding a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1 or a polynucleotide comprising a naturally occurring polynucleotide sequence at least 70% identical to a polynucleotide sequence of SEQ ID NO:2, using well known methods of sequence analysis without any undue experimentation. For example, the identification of relevant polynucleotides could be performed by hybridization and/or PCR techniques that were well-known to those skilled in the art at the time the subject application was filed and/or described throughout the Specification of the instant application. See, e.g., page 12, line 13 through page 13, line 9; page 25, lines 2-6 and 18-28; and Example VI at pages 45-46. Thus, one skilled in the art need not make and test vast numbers of polynucleotides. Instead, one skilled in the art need only screen a cDNA library or use appropriate PCR conditions to identify relevant polynucleotides that already exist in nature. The skilled artisan would also know how to use the claimed polynucleotides, for example in expression profiling, disease diagnosis, or detection of related sequences as discussed above.

The specification also describes the expression vectors into which the claimed fragments could be inserted, and the construction of fusion proteins (pages 22-24 and page 47, line 8 through page 48, line 3). The specification describes, for example, specific assays for myosin activity on page 48; binding assays to detect molecular interactions of “MHCH or biologically active fragments thereof” on page 50, lines 4-19; and immunological methods for detecting and measuring MHCH on page 25, lines 7-16. These methods could be used to detect and characterize peptide variants and fragments of SEQ ID NO:1. Given this guidance, one of ordinary skill in the art would readily understand how to select and screen polynucleotides encoding fragments of SEQ ID NO:1 with ATPase activity or immunogenic activity without any undue experimentation.

Furthermore, the claims are directed to polynucleotides, not polypeptides, and it is the functionality of the claimed polynucleotides, not the polypeptides encoded by them, that is relevant. Members of the claimed genus of variants may include, for example, mutant alleles associated with diseases, or single nucleotide polymorphisms (SNPs). Members of the claimed genus of variants may be useful even if they encode defective MHCH polypeptides. For example, the variant polynucleotides could be used for the detection of sequences related to MHCH (see the specification at page 25, lines 17-28, and page 36, lines 24-30) including MHCH variants that may be associated with disease states, such as the diseases listed on page 27, line 16 through page 28, line 3, of the specification. See the specification at, for example, pages 36-40 for disclosure of how to use the claimed sequences in diagnostic assays.

The Examiner has cited Attwood et al., identifying some of the difficulties that may be involved in predicting protein function; however, this reference does not suggest that functional homology cannot be inferred by a reasonable probability in this case. At most, this article suggests that it is difficult to make predictions about function with certainty. The standard applicable in this case is not proof to certainty, but rather, proof to a reasonable probability. In fact, Attwood et al. point out the value of sequence analysis, in particular with regard to the identification of conserved motifs in proteins. “Because motifs usually reflect some vital structural or functional role (Fig. 2), they **effectively** provide diagnostic family signatures” (Emphasis added; Attwood et al. p. 332, col. 2).

An analysis of the sequence of SEQ ID NO:1 shows that it contains conserved residues and structural motifs characteristic of members of the myosin family. For example, the specification shows alignments of SEQ ID NO:1 with *C. elegans* myosin I heavy chain and *H. annuus* unconventional myosin heavy chain, and points out regions of homology and conserved amino acid residues in the three proteins (Specification at Figure 2). The specification, on page 17, line 26 through page 18, line 9, identifies specific residues and myosin signatures within SEQ ID NO:1, including the myosin head domain, which is known to contain the ATPase activity and actin binding sites in myosin motor proteins. The Examiner’s attention is directed to Exhibit A which shows the identification of the myosin motor head domain in SEQ ID NO:1 by HMMER analysis of the PFAM database. At the time of filing of the instant application, the crystallographic structure of a myosin motor head was available to assist one of skill in the art in the determination of “specific catalytic residues and structural motifs,” particularly

those critical for ATPase activity and actin binding (See the enclosed reference of Rayment et al. (1993) Science 261:50-58).

Further, the Examiner requires working examples (Office Action, page 4). There is no such requirement under the law to provide "working examples." As set forth in *In re Borkowski*, 164 USPQ 642, 645 (CCPA 1970) (footnote omitted):

However, as we have stated in a number of opinions, a specification need not contain a working example if the invention is otherwise disclosed in such a manner that one skilled in the art will be able to practice it without an undue amount of experimentation.

See also M.P.E.P. 2164.02 as follows:

Compliance with the enablement requirement of 35 U.S.C. 112, first paragraph, does not turn on whether an example is disclosed. An example may be "working" or "prophetic"... A prophetic example describes an embodiment of the invention based on predicted results rather than work actually conducted or results actually achieved.

Thus, there is no requirement under the law to provide "working examples" of what is claimed. Rather, one looks to whether the specification provides a description of how to make what is claimed. The present specification provides the requisite description.

Contrary to the standard set forth in *Marzocchi* and *Borkowski*, the Examiner has failed to provide any *reasons* why one would doubt that the guidance provided by the present specification would enable one to make and use the recited polynucleotides. Hence, a *prima facie* case for non-enablement has not been established. For at least the above reasons, withdrawal of the enablement rejections under 35 U.S.C. § 112, first paragraph, is respectfully requested.

Written description rejections under 35 U.S.C. § 112, first paragraph

Original claims 3-6 and 9-14, now replaced by new claims 23-31 have been rejected under the first paragraph of 35 U.S.C. 112 for alleged lack of an adequate written description. This rejection is respectfully traversed.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession of *the invention*.

The invention is, for purposes of the “written description” inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office’s own “Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1”, published January 5, 2001, which provide that :

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics⁴² which provide evidence that applicant was in possession of the claimed invention,⁴³ i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics.⁴⁴ What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail.⁴⁵ If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met.⁴⁶

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:2 are specifically disclosed in the application (see, for example, page 17, lines 19-34). Variants of SEQ ID NO:1 and SEQ ID NO:2 are described, for example, at page 18, lines 18-33. Incyte clones in which the nucleic acids encoding the human myosin heavy chain homolog were first identified and libraries from which those clones were isolated are described, for example, at page 17, lines 19-25 of the Specification. Chemical and structural features of SEQ ID NO:1 are described, for example, on page 17, line 26 through page 18, line 9. Given SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1 having 90% sequence identity to SEQ ID NO:1. Given SEQ ID NO:2, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:2 having 70% sequence identity to SEQ ID NO:2. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

A. The Specification provides an adequate written description of the claimed variants and fragments of SEQ ID NO:2.

The Office Action has further asserted that the claims are not supported by an adequate written description because the specification “only provides the following representative species encompassed by these claims: a polynucleotide consisting of the nucleotide sequence of SEQ ID NO:2 and a polynucleotide encoding a polypeptide consisting of the amino acid sequence of SEQ ID NO:1.... Given this lack of additional representative species as encompassed by the claims, Applicants have failed to sufficiently describe the claimed invention”...

Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which “DNA claims” have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

In claims to genetic material, however, a generic statement such as “vertebrate insulin cDNA” or “mammalian insulin cDNA,” without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count: A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, “an mRNA of a vertebrate, which mRNA encodes insulin” in *Lilly*, and “DNA which codes for a human fibroblast interferon-beta polypeptide” in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides in terms of chemical structure, rather than on functional characteristics. For example, the “variant language” of independent claim 30 recites chemical structure to define the claimed genus:

30. An isolated polynucleotide selected from the group consisting of:...
- b) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 70% identical to a polynucleotide sequence of SEQ ID NO:2...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:2. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides. The polynucleotides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry “on

whatever is now claimed,” the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*

2. The present claims do not define a genus which is “highly variant”

Furthermore, the claims at issue do not describe a genus which could be characterized as “highly variant.” Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner’s attention is directed to the enclosed reference by Brenner et al. (“Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships,” Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to myosin proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as myosin proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The “variant language” of the present claims recites, for example, a polynucleotide encoding “a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1” and “a polynucleotide comprising a naturally occurring polynucleotide sequence at least 70% identical to a polynucleotide sequence of SEQ ID NO:2” (note that SEQ ID NO:1 has 612 amino acid residues). This variation is far less than that of all potential myosin proteins related to SEQ ID NO:1, i.e., those myosin proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The ‘525 patent claimed the benefit of

priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the “dark ages” of recombinant DNA technology.

The present application has a priority date of November 5, 1998. Much has happened in the development of recombinant DNA technology in the 19 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1 and SEQ ID NO:2, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

4. Summary

The Office Action failed to base its written description inquiry “on whatever is now claimed.” Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1 or SEQ ID NO:2. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides defined by the present claims is adequately described, as evidenced by Brenner et al and consideration of the claims of the ‘740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

Rejection under 35 U.S.C. § 112, second paragraph

Claims 4 and 5 have been rejected under 35 U.S.C. § 112, second paragraph, as allegedly being “indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention” (Office Action, page 6). Claim 5 has been canceled. Therefore, the rejection with respect to this claim is moot.

Original claim 4 was allegedly indefinite because “the specific nucleotide sequence of the polynucleotide to which the claimed polynucleotide has 70% identity is not known and not stated in the claim.” New claims 23 and 30, now replace claim 4. Claim 23 recites an isolated polynucleotide encoding a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:1. Claim 30 recites an isolated polynucleotide comprising a naturally occurring polynucleotide sequence at least 70% identical to a polynucleotide sequence of SEQ ID NO:2. Given the sequences of SEQ ID NO:1 and SEQ ID NO:2, which are disclosed in the instant application, one of skill in the art could readily understand the scope of the claimed invention. Therefore, withdrawal of the rejection under 35 U.S.C. § 112, second paragraph is respectfully requested.

Rejection under 35 U.S.C. § 102

Original claims 3 and 9, now replaced by claims 23 and 31, are rejected under 35 U.S.C. § 102 as allegedly being anticipated by the references of Calabretta et al. (U.S. Patent No. 5,734,039) and Dahlberg et al. (U.S. Patent No. 5,541,311) on the grounds that the references teach the claimed polynucleotide fragments.

As currently pending, claim 23 recites a polynucleotide encoding a biologically active fragment of a polypeptide having an amino acid sequence of SEQ ID NO:1, wherein said fragment has ATPase activity, and a polynucleotide encoding an immunogenic fragment of a polypeptide consisting of an amino acid sequence of SEQ ID NO:1, wherein said fragment comprises **at least 15** contiguous amino acid residues of SEQ ID NO:1. Claim 31 recites an isolated polynucleotide consisting of **at least 25** contiguous nucleotides of SEQ ID NO:2, or the complement thereof. Support for the new claims can be found in the specification, for example, at page 8, lines 36-41, which defines the term “fragment,” page 17, line 26, through page 18, line 9, which describes the homology between MHCH and myosin, and at page 48, lines 5-22, which describes assays for myosin ATPase activity.

The polynucleotide sequence disclosed by the Calabretta reference does not encode a polypeptide containing 15 contiguous amino acid residues of SEQ ID NO:1, nor a biologically active fragment of SEQ ID NO:1 having ATPase activity. The polynucleotide sequence disclosed by the reference of Dahlberg et al. does not contain 25 contiguous nucleotides of SEQ ID NO:2. Therefore, the references do not disclose the claimed polynucleotide fragments, and Applicants respectfully request withdrawal of the rejections under 35 U.S.C. § 102.

CONCLUSION

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact Applicants' Attorney at (650) 855-0555.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. 09-0108.

Respectfully submitted,
INCYTE CORPORATION

Date: July 7, 2003

Jenny Buchbinder
Jenny Buchbinder
Reg. No. 48,588
Direct Dial Telephone: (650) 843-7212

Date: July 7, 2003

Cathleen M. Rocco, Reg No 41,113
For Cathleen M. Rocco
Reg. No. 46,172
Direct Dial Telephone: (650) 845-4587

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE:
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wyntford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate

on of specific and cancer or previous

zymes (including
ible by drugs and
ng transcriptional
additional cellular

Accordingly, the
plement of genes
ne development of
enzyme induction
chemical-induced
verse reactions to
some of which are
l phenomenon *per*
filing technologies
ols in target tissues
sms of xenobiotic-
in target tissues is
generated in the
ntification of toxic
and contributing
if the gene profile
ed *in vivo* could be
ve of all new drug
f toxicity, thereby
of such toxicants.
onality of all genes
r term goal, as the
their functionality
a *pattern* of gene
hed to that of well-
in vivo similarities
platform for more
beginning to gain
ercially producing
city assessment of
some of which are
ally-related to any
in broad-spectrum
ne arrays are now
es in growth factor
hemically-induced

anges presents a
of development and
npting this difficult-
methods have been
s that demonstrate

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinanuer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Duguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

Differential gene expression

659

sequences from
ed from normal.
lter replicas are
herwise) labelled
NA populations.
ll population will
the treated cells.
be used to probe
hich are only up-
s (1979) screened
n order to obtain
ndbreaking in its
ning, as up to 2
are differentially
ent way to check
ompleted.

e success of early
on gave rise to a
be developed was
low). In general,
opulation (tester)
eparation of the
ybridized common
l through the use

ysical separation
Several methods
chromatography
d Dinauer 1990)
proach, common
om control cells)
phy, as hydroxy-
sorbed cDNA is
entially expressed
tly as a probe to
1984, Hedrick *et al.*

sitivity enhancing
of the problems
guid and Dinauer
ystems as a means
the control and
or ('oligovec r',

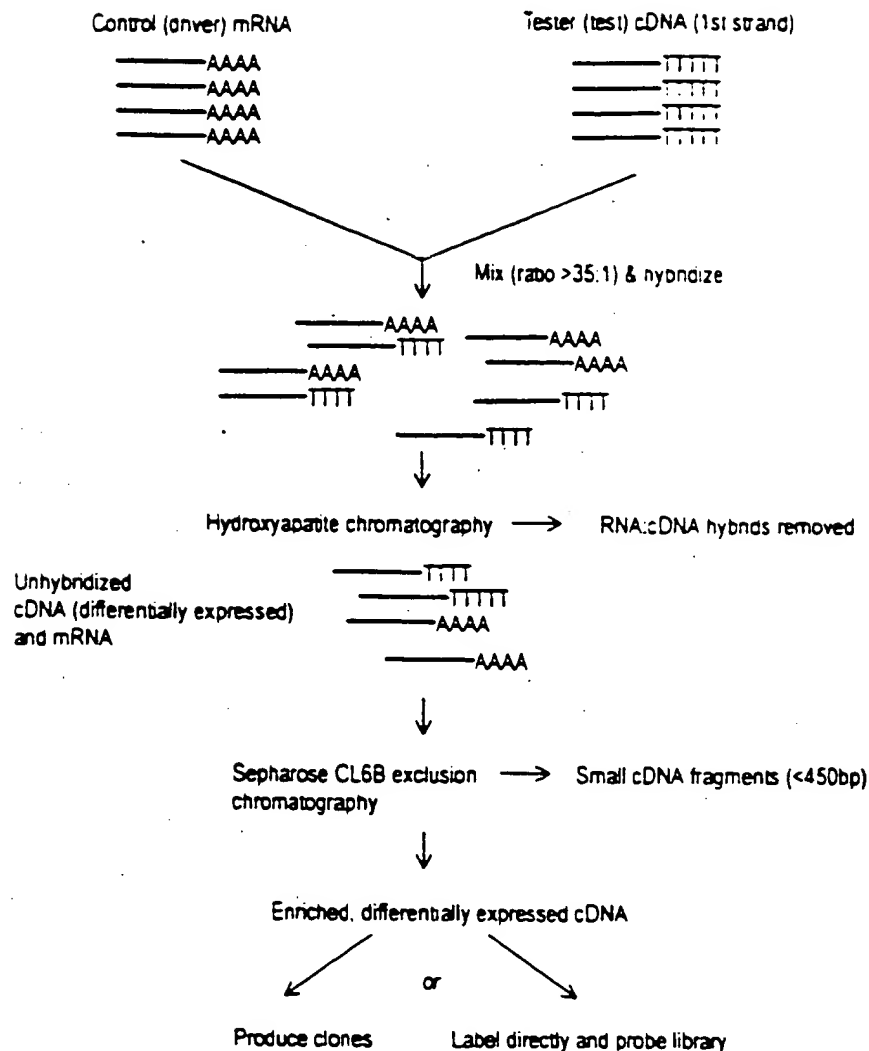


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/ altered (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/ altered population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

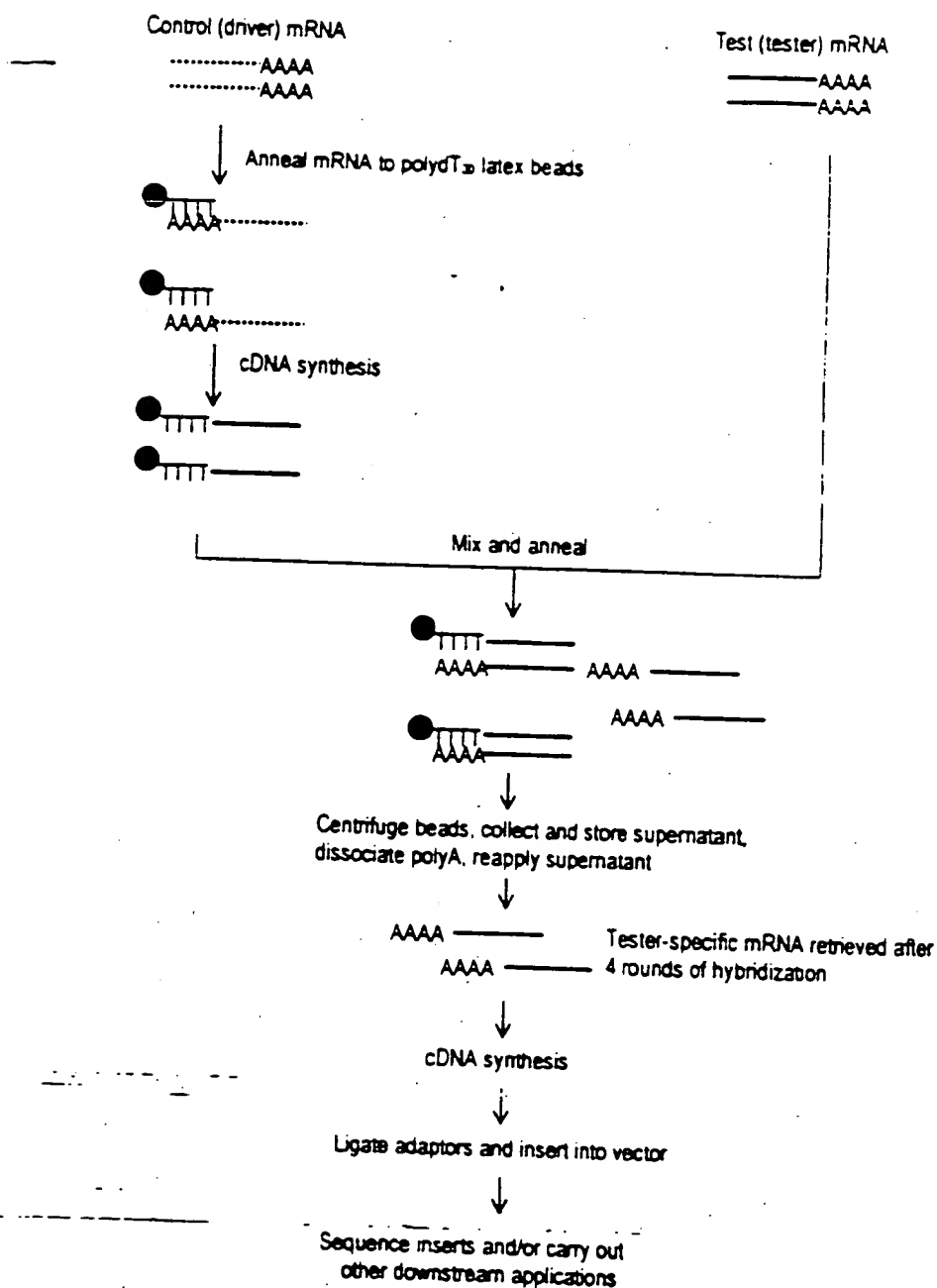


Figure 2. The use of oligodT₂₀ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/alterd (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara et al. (1994).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT)₃₀ primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT₃₀ forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT₃₀ population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promoter sequence

mRNA
AA
AA

removed after

A extracted from the
dT oligonucleotides
isolation is repeatedly
isolation of mRNA is
stream applications, as

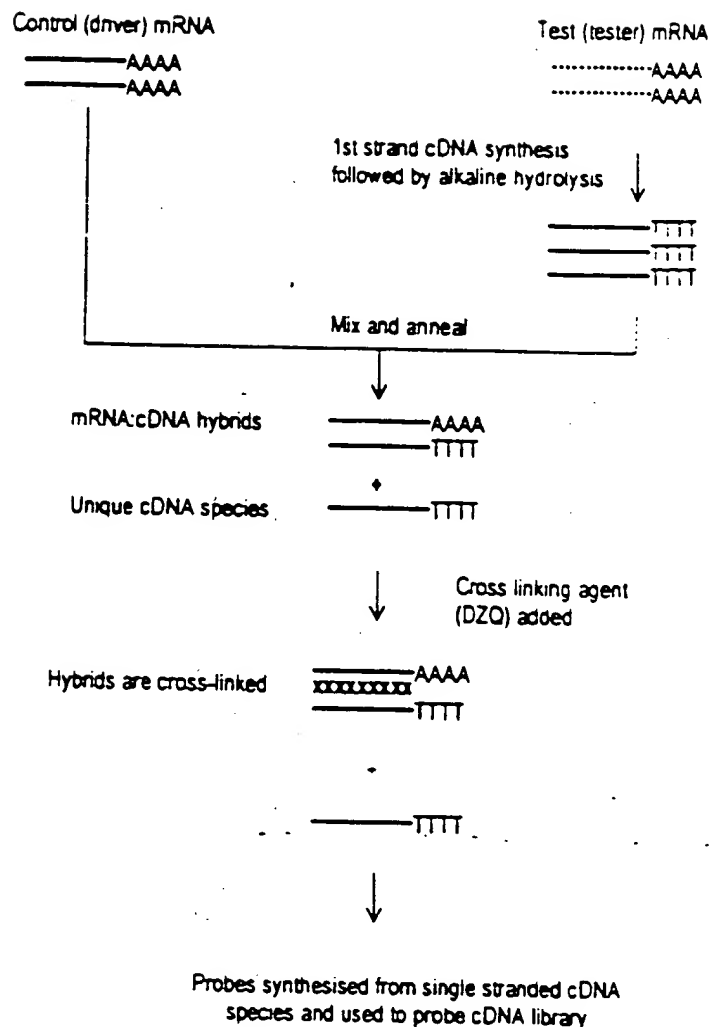


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,2-diazobenzoyl-1,4-benzoquinone (DZO) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/ μ g total RNA
Abundant	12 000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11 000	0.004	0.002

Modified from Bertoli *et al.* (1995).

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

mRNA
AAAA
AAAA

TTTT
TTTT
TTTT

with 1st strand tester
cross linked with 2.5
tes are differentially
using Sequenase 2.0
oes not react with the
ed to screen a cDNA
er *et al.* (1996), with

alian cell.

mass
each
9/μg
RNA

5
4
02

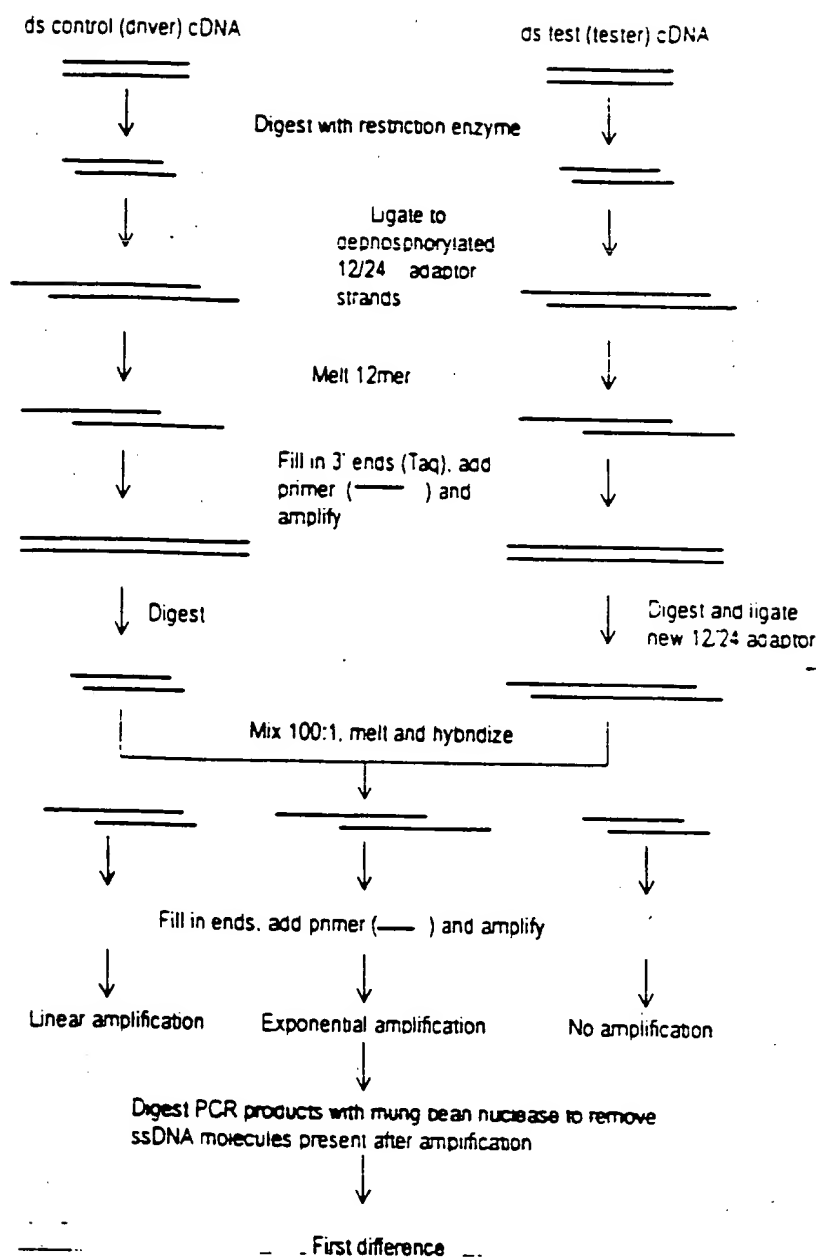


Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *DpnII*. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3' ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *DpnII*. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complementary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

and tester cDNA are
12/24 adaptor strands
products. The 12mer is
reverse transcriptase. Each cDNA
adaptor is removed with
amplified tester cDNA
of driver. The 12mer
with primers identical
which are exponentially
cDNA products are
This is digested and a
from the hybridization
described by Lisitsyn *et al.*

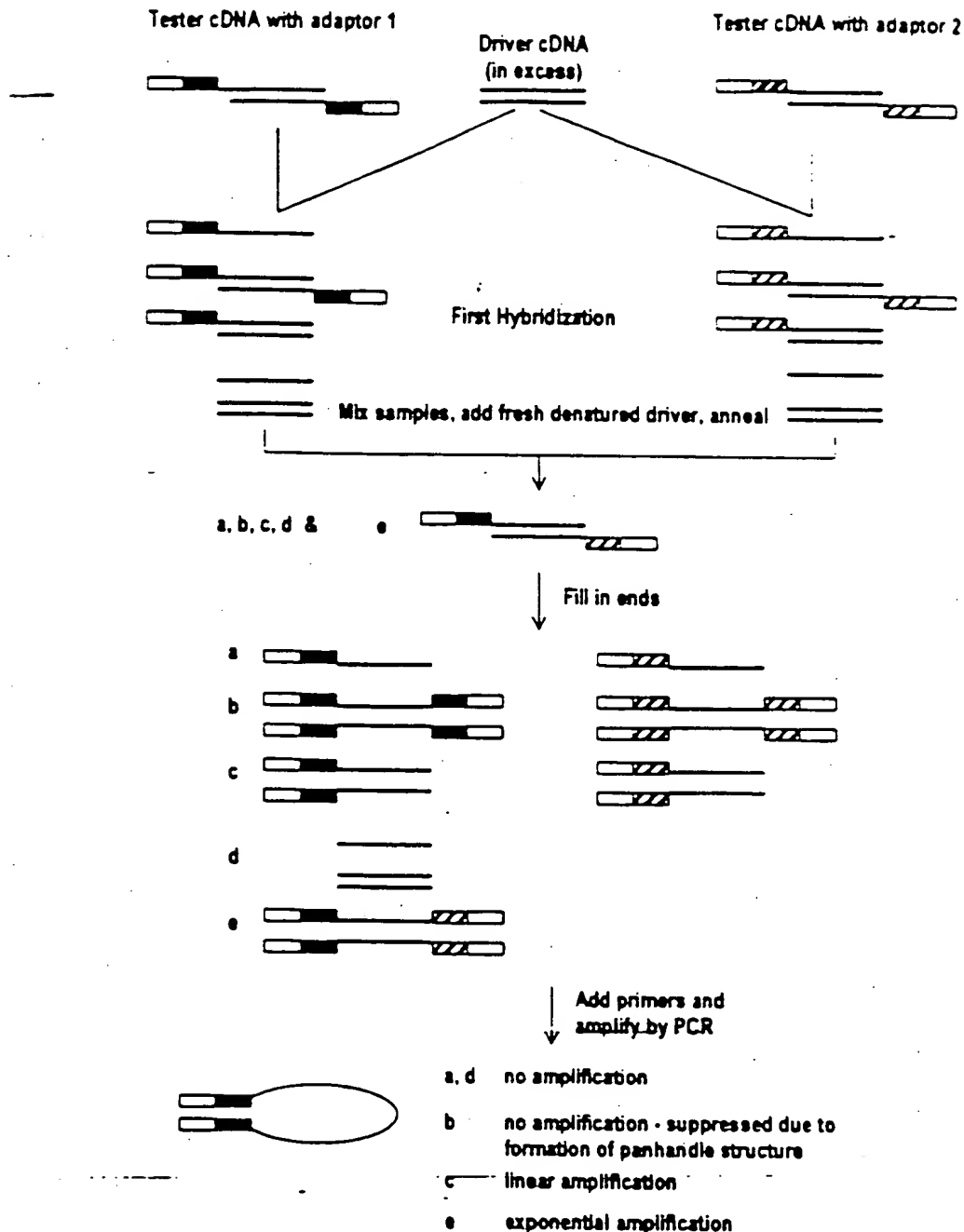


Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurakaya *et al.* (1996), with permission.

Differential gene expression

267

cDNA with adaptor 2

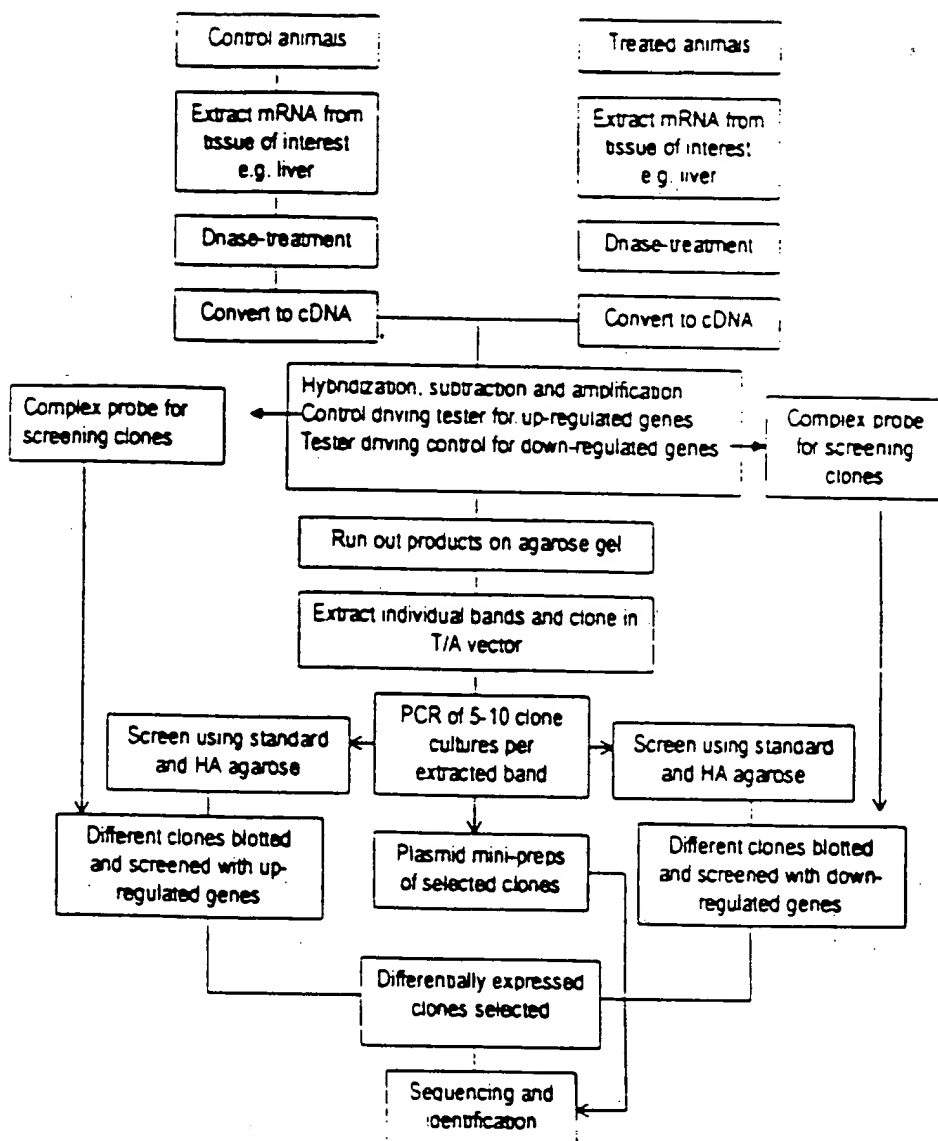
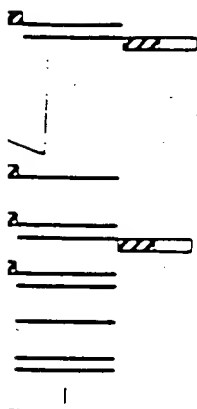


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

ccess of driver cDNA is and allowed to hybridize abundant molecules; and t differentially expressed ation, the two primary driver can also be added ences. Type e molecules ified using two rounds of directly or cloned into a il. (1996) and Gurakaya

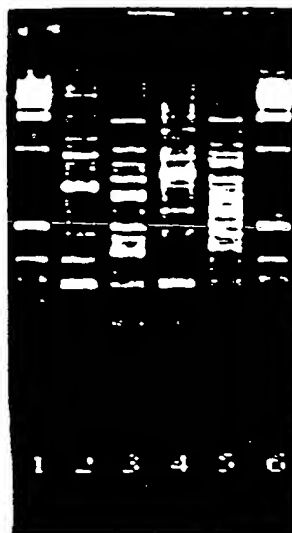


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with Wy-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy-14,643 treatment; 3—genes downregulated following Wy-14,643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy-14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete functional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

Differential gene expression

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1-4, 6, 9, 13, 14, and 17-20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopexin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares pJNMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares pJNMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4-6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

sent with WY-14,643 or
s used to generate the
lontech). Lane: 1-1kb
down-regulated following
al treatment; 5-genes
roduced from Rockett *et*

ained. For example,
gen Wy-14,643, up-
the rat (a sensitive
n the guinea pig, a
shed observations).
wn-regulated in the
named TAPA-1) is
e number of cellular
erentiation (Levy *et*
t in the phenomena
is intriguing, and
erentially regulated
of this approach is
base sequences, but
enes of completely
l assessment of the
he lack of complete
ie profiling studies
nbiotic challenge,
r further detailed

rimed PCR' (Liang
d to as 'differential

display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' (dT₁₁)CA 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (arbitrary primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

- (1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).
- (2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.
- (3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cyt plasmic RNA rather than total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

may be recovered for
ed with which it can
to make and identify

hods of priming the
th a 2-base 'anchor'
)'. Alternatively, an
(Welsh *et al.* 1992).
' (RNA Arbitrarily
CR products may be
imes. In addition, it
y bacterial mRNAs
e transcription and
an arbitrary primer
ompared to *random*
ion). The resulting
the system (primer
ly includes 50-100
ination of different
ecies from a cell can
lations are analysed
an be identified and
is.

oday for identifying
ved disadvantages:

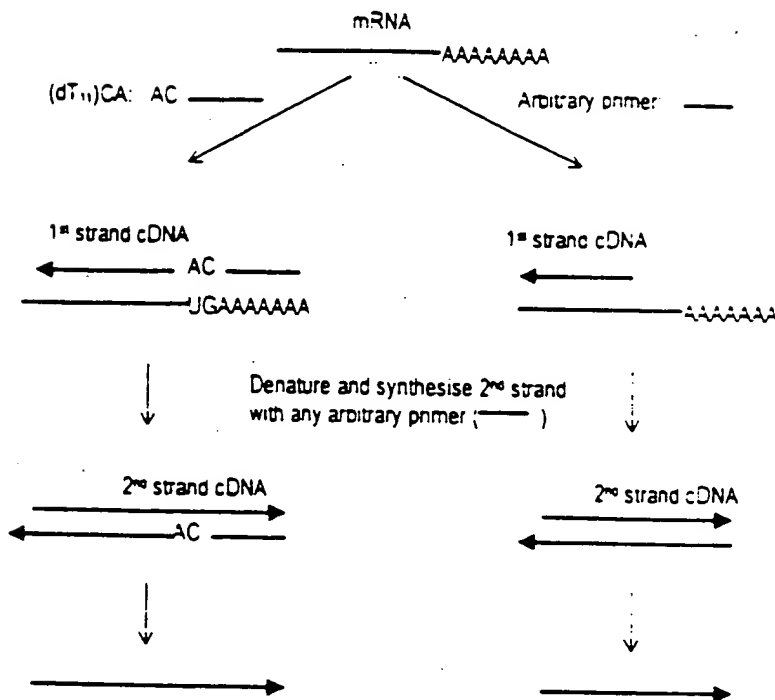
NAs (Bertioli *et al.*
the isolation of very
nces (Guimeraes *et*

end of the mRNA
always be the case
ed in Genbank and
D cannot always be

y often cannot be
up to 70% of cases
uce false positives,
d Denman 1997),
e (Burn *et al.* 1994)
d and two induced
ted that the use of
tives arising fr m

esses of the DD
(1996) and from

Differential gene expression



cDNA can now be amplified by PCR using original primer pair

Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT₁₁NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

split into two and different adaptors ligated to the 5' ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5' digestion products. The use of restricted 3'-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3' cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobilized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

Differential gene expression

673

group. Incorporated
restriction enzyme—one
cognition sequence.
with the IIS enzyme.

released. The two
amplified products are
dimers are formed in
hundreds of gene tags
the number of times
percent of that gene's
genes identification of

cal difficulty of the
d towards abundant
mic setting and has
e.

ating differentially
sky (1995). In this
(dT) primers. The
and captured with
wanted 5' digestion
the complexity of the
es is represented by
acilitate subsequent
with one adaptor-
fied population is
the dissociation. The
nt adaptor-specific
mobilized 3' cDNA
tion endonucleases
result is a fingerprint
ential digests used).
entify differentially
l and cloned. The
roducible, and the
olved in the final
rarely resolve more
r more which are
of 2-D gels such as
(1991) may help t

ragments was later
ead of sequential
ese authors simply
is without further

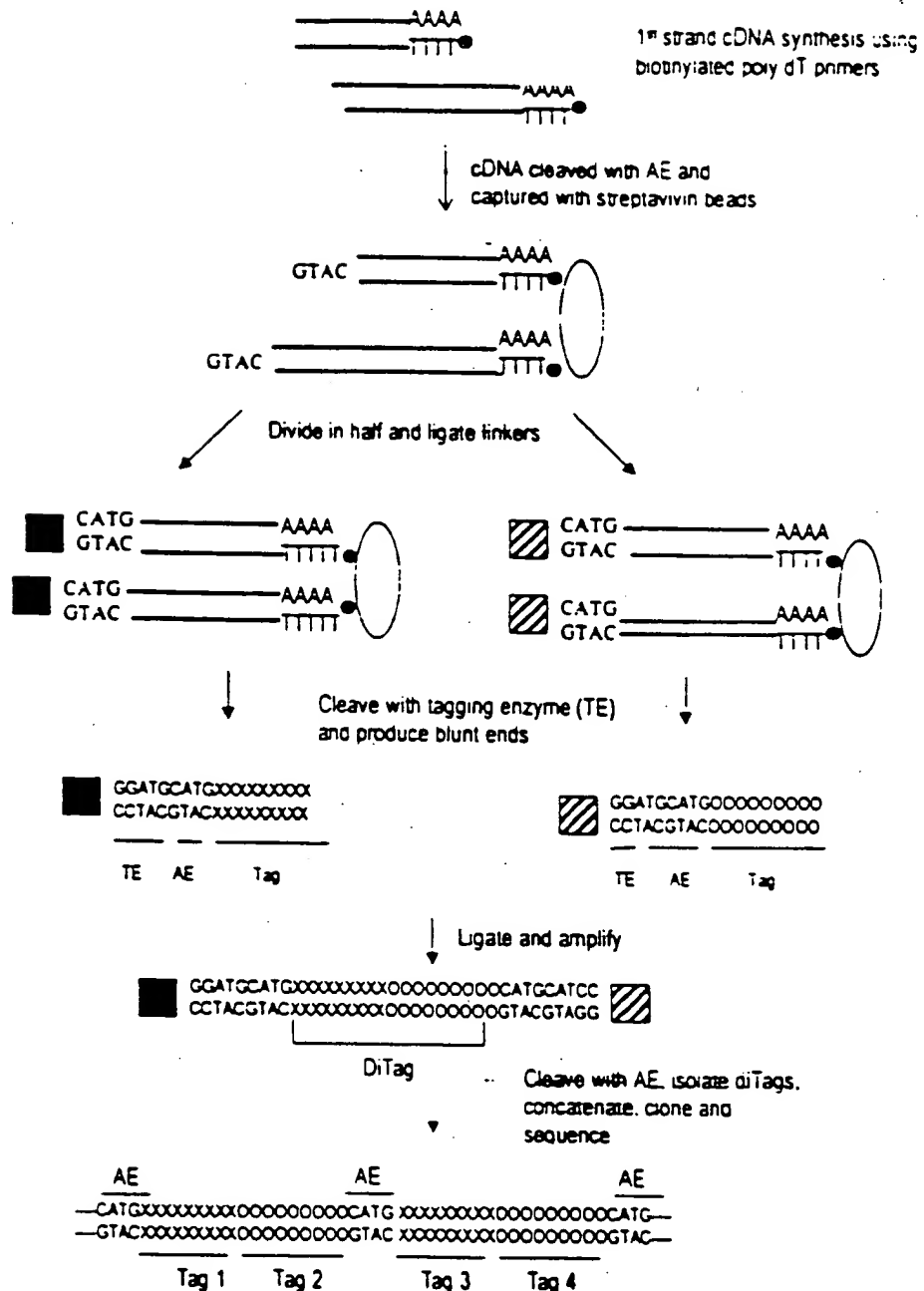


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the diTags isolated from the linkers using PAGE. The diTags are then ligated (during which process, concatenation occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

it takes a great deal that they are indeed issue. Normally, the R. Even so, each of al of rapid analysis the development of Schena *et al.* 1996), ial gene expression 'chips' containing le copies of part of proven involvement : cellular processes. and animal species. id a few companies h Laboratories and or even thousands DNA from the test hen analysed, with antitative means to is. Of course, there ch are in the array to elucidating the ent system may be rectly identify and ions, and an open ntially expressed. r of gene fragments ed gridding up to high density chip- duced off-the-shelf d determination of . Aside from their and probing DNA newer micro-array roducible between resolved within the

ssed genes nes obtained from identity (putative rapid and efficient : profiles of gene- dams *et al.* (1991), ated that there are resenting over half

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatazis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatazis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromised tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15 000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30 000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005 % of the total population (table 1). Bertoli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

998, Kas-Deelen *et al.* 1998).

issue unimportant. After all, since all cell type could in molecular mechanisms growth. It is perhaps in *vitro* as in *vivo* cells probably molecular changes that

biological variation is being used. It is not ways to identical isoquine oxidation and determines the (1993, Meyer and complex, but allelic and mental health es. Careful thought and to the possible effect of this can be unimportant minor lual animals, thus mechanisms of the may be of utmost amb to or resist the

high percentage of

suggesting that mammals at any one time although figures as edrick *et al.* (1984) the rare abundance table 1.

been compared with not all differentially ular, rare messages or easily recovered, as the majority of opulation (table 1). tes (heterogene us ble to detect mRNA

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10 000 × smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- γ -stimulated HeLa cells differentially express up to 433 genes (assuming 24 000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38 000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockert, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3' polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

y effective—proving numbers of artificial are messages already odels will genuinely n addition, there are mRNAs may have plification by PCR-cumstances not all lopment, deadenyl- iteitz 1998), whilst sp70 (and perhaps, alle *et al.* 1994). The ficiency of systems of any system also display techniques olate mRNA that is : used to prime first ed to some degree has been shown, at a lead to inefficient ibration kit user kewise in other SH n amplification step sequences amplify

temporal factor. It interrogate a cell at nes showing altered ease processes and ides of signalling, which are switched al information may ormation about the can be derived for iar interest to the me point analysis is h, of course, adds

ie of how large the ne in question with isolation of genes ported using SSH trating a change in e is a 'grey zone' isolation between

experiments and animals. DD, on the other hand, is not subject to this 'grey zone' since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaP r HR (National Diagnostics, Hesse, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

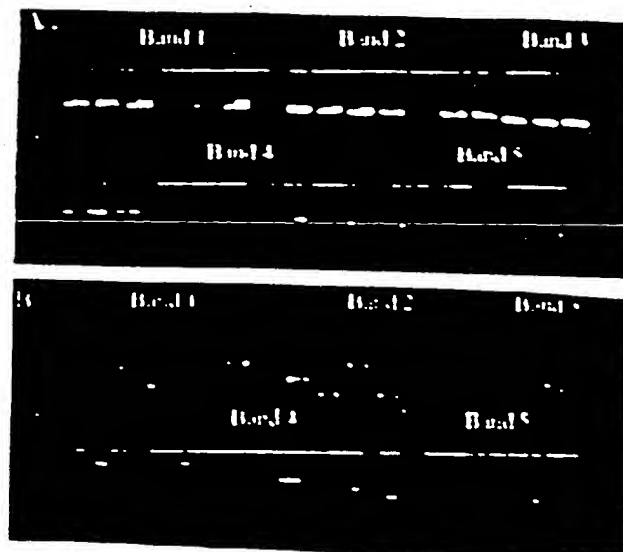


Figure 10. Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1993) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, over staining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to over stain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000+ bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

HA-red. Bands of decreasing subtractive hybridization each cloned band and their high resolution 2% agarose gel. With few exceptions, all bands contain the presence of HA-red. The percentage of GC within each band. For the same size, at least four

each gel should indicate and separate otherwise identical content. Geisinger identifying DD-derived bands in laboratory on clones

the differential display method out in a standard method and incorporated

there being different GC content. However, again, one might use GGE) or temperature contents of a band, or on the reamplified

ies to visualize large problem in that, in terms of bands. One approach to be used by Litterlinden *et*

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitimate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 or its as yet undiscovered sister X_2 ? For example, using SSH, part of a gene was isolated,

ne altered expression primers and/or post-receptors, cell cycling considered as candidates rays (e.g. Clontech's is to some degree by apoptosis, stress, DNA-

length amongst the *et al.* 1994, Sun *et al.* ves varies with the tors which have not through illegitimate y can arise through false positives appear some may arise from or technical reasons. es can be carried out a probes synthesized d clones (Hedrick *et al.* es will hybridize to h is that rare species for those using SSH e subtracted cDNA reverse subtraction ches rare sequences, ating low abundance d to go back to the a more quantitative ts, the sensitivity is ods for accurate and

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96+ -well plates and

l products which are ly reduce the size f leads to a reduced mbers whose DNA e.g. the cytochrome e identified as being other gene X₁ or its a gene was isolated,

appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN- γ , Frye *et al.* 1989), β -actin (Heuvel *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β -2-microglobulin (β -2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using the differential display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

ve analysis is more internal standard, the e is often excessive, s of gene species. The ntively involved. One ange in the test cells r tried in the past, for (Heuval *et al.* 1994), ng *et al.* 1994), di- microglobulin (β -2- rase (HPRT, Foss *et* Ideally, an internal egardless of cell age, However, it has been ping genes currently in conditions and in therefore, that pre- ng genes to establish

d with caution. By ion one can perhaps s to external stimuli. oxic effects of a wide uinea pigs are largely ke *et al.* 1989, 1993, e reason(s) why is to ntify those which are edge of the effects of oxic carcinogenesis complex. Perhaps if oxic effects and it was -regulated five times he importance of the change in expression example, what is the particular treatment, es the literature one up-regulated 40-60- -fold increase would ene Z has never been a makes your 5-fold s if that same 5-fold eatment with related

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5-10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

an easily obtainable test animals/cells in uli. However, it has

carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as described extensively herein, but rather protein-protein, protein-DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

ally Darney and Chris
manuscript prior t
with the policy of the

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPoulos, M. H., XIAO, H., MERRILL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., McCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.

AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342-346.

AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247-254.

BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249-1255.

BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.

BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520-4523.

BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37-46.

BURN, T. C., PETROVICK, M. S., HOMAU, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776-2783.

CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447-451.

CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13-20.

CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445-4452.

CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YACER, J. D., 1996, Identification of genes whose expression is altered during mitosisuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783-2786.

CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793-794.

CLONTECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18-19.

CLONTECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15-16.

DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194-2198.

DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646-3659.

—, DERISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.

DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025-6030.

DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1-9.

DUGUID, J. R. and DINALE, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789-2792.

DUNBAR, P. R., OCC, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDULO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413-416.

- FITZPATRICK, D. R., GERMAIN-LEE, E. and VALLE, D., 1995. Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457-466.
- Foss, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998. Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, **9**, 67-78.
- FRYE, R. A., BENZ, C. C. and LIU, E., 1989. Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153-1157.
- GEISINGER, A., RODRIGUEZ, R., ROMERO, V. and WETTSTEIN, R., 1997. A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, <http://no.trends.com>, document T01110.
- GRAESS, T. M., HOMMEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992. Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609-619.
- GRIFFIN, G. and KRISHNA, S., 1998. Cytokines in infectious diseases. *Journal of the Royal College of Physicians*, **32**, 195-198.
- GROENINK, M. and LEIGWATER, A. C. J., 1996. Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechiques*, **XI**, 23-24.
- GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and McCLANAHAN, T., 1995b. A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, **121**, 3335-3346.
- GUIMARAES, M. J., LEE, F., ZLOTNIK, A. and McCLANAHAN, T., 1995a. Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, **23**, 1832-1833.
- GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAYEV, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996. Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-Myristate 13-Acetate. *Analytical Biochemistry*, **240**, 90-97.
- HAMPSON, I. N. and HAMPSON, L., 1997. CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22-24.
- HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996. Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, **24**, 4832-4835.
- HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992. Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.
- HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991. Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097-7104.
- HATADA, I., HAYASHIZAKI, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (U.S.A.)*, **88**, 9523-9527.
- HECHT, N., 1998. Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555-561.
- HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984. Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149-153.
- HERTZ, R., SECKBACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996. Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218-224.
- HEUVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994. Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transcriptase-polymerase chain reaction. *Cancer Research*, **54**, 62-68.
- HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPPELLI, B., CHISSOR, S., DIETRICH, N., DUBROU, T., FAYELLO, A., GISH, W., HAWKINS, M., HUTTMAN, M., KUCERA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R. and MARRA, M., 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807-828.
- HUBANK, M. and SCHATZ, D. G., 1994. Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640-5648.
- HUNTER, T., 1991. Cooperation between oncogenes. *Cell*, **64**, 249-270.
- IVANOVA, N. B. and BELYAVSKY, A. V., 1995. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954-2958.
- JAMES, B. D. and HIGGINS, S. J., 1985. *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).
- KAS-DEZLEN, A. M., HARMSSEN, M. C., DE MAAR, E. F. and VAN SON, W. J., 1998. A sensitive method for

- characterisation of rat and hydratase. *Genomics*, 27.
- f hypoxanthine phospho-a-actin mRNA expression by differential polymerase
- apic method for screening *Elsevier Trends Journals*
- I. H., 1992. Hybridisation derived from whole tissues.
- ial of the Royal College of
- enes associated with liver techniques. *XI*, 23-24.
- DI, J. C., LEE, F. and ic development in the yolk
- , Differential display by 2-1833.
- PEKOV, G. L., LUKYANOV, SVERDLOV, E. D., 1996. Polymerase chain reaction: phorbol 12-Myristate 13-
- ng made easy. *Life Science*
- m oligonucleotide primed e cDNA cloning. *Nucleic*
- ul cross linking subtraction ion probes. *Nucleic Acids*
- ive cDNA cloning using undifferentiated human
- AI, T., 1991. A genomic marks. *Proceedings of the*
- Bioessays*, 20, 555-561.
- olation of T cell-specific
- ononal suppression of the *Biological Chemistry*, 271.
- W. F., LUCIER, G. W. and sponse relationships using research, 54, 62-68.
- CHISSOE, S., DIETRICH, N., UCABA, T., LACY, M., LE, C., RIFKIN, L., ROHLFING, EVASKIS, E., UNDERWOOD, S. Generation and analysis 328.
- ession by representational
- ressed genes by restriction arch, 23, 2954-2958. RL Press Ltd).
98. A sensitive method for quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, 5, 622-626.
- KILTY, I. and VICKERS, P., 1997. Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter (Stratagene)*, 10, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998. Position effect in human genetic disease. *Human and Molecular Genetics*, 7, 1611-1618.
- KO, M. S., 1990. An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, 18, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993. Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, 101, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989. Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, 99, 148-160.
- LENNARD, M. S., 1993. Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, 9, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998. CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, 16, 89-109.
- LIANG, P. and PARDEE, A. B., 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992. Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, 52, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993. Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, 21, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995. Analysis of altered gene expression by differential display. *Methods in Enzymology*, 254, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FLUNK, W. D. and VILLEPONTEAU, B., 1995. Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, 23, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WICLER, M., 1993. Cloning the differences between two complex genomes. *Science*, 259, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995. REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, 18, 200-202.
- LUNNEY, J. K., 1998. Cytokines orchestrating the immune response. *Reviews in Science and Technology*, 17, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992. Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, 7, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998. The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, 53, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996. Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, 24, 1504-1507.
- McKENZIE, D. and DRAKE, D., 1997. Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter (Stratagene)*, 10, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996. RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, 11, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981. Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, 88, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997. Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, 37, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991. Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, 28, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TRIO FOJO, A. and BATES, S. E., 1990. Use of the polymerase chain reaction in the quantitation of the *mdr-1* gene expression. *Biochemistry*, 29, 10351-10356.
- NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYERISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996. Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, 6, 1-42.

- NISHIO, Y., AIELLO, L. P. and KING, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103-106.
- O'NEILL, M. J. and SINCLAIR, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681-2682.
- ORTON, T. C., ADAM, H. K., BENTLEY, M., HOLLOWAY, B. and TUCKER, M. J., 1984, Clobazam: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138-151.
- PELKONEN, O., MAENPAA, J., TAAVITSAINEN, P., RAUTIO, A. and RAUNIO, H., 1998, Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, **28**, 1203-1253.
- PHILIPS, S. M., BENDALL, A. J. and RAMSHAW, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199-203.
- PRASHAR, Y. and WEISSMAN, S. M., 1996, Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, **93**, 659-663.
- RAGNO, S., ESTRADA, I., BUTLER, R. and COLSTON, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143-146.
- RAMANA, K. V. and KOHLI, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437-446.
- RICHARD, L., VELASCO, P. and DETMAR, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1-6.
- ROCKETT, J. C., ESDAILE, D. J. and GIBSON, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, **22**, 329-333.
- RODRICKS, J. V. and TURNBULL, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197-212.
- ROGLER, G., HAUSMANN, M., VOGL, D., ASCHENBRENNER, E., ANDUS, T., FALK, W., ANDRESEN, R., SCHOLMERICH, J. and GROSS, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205-215.
- ROHN, W. M., LEE, Y. J. and BENVENISTE, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311-330.
- RUDIN, C. M. and THOMPSON, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435-446.
- SAKAGUCHI, N., BERGER, C. N. and MELCHERS, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139-2147.
- SAMBROOK, J., FRITSCH, E. F. and MANIATIS, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Ferguson (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6-37.
- SARGENT, T. D. and DAWID, I. B., 1983, Differential gene expression in the gastrula of *Xenopus laevis*. *Science*, **222**, 135-139.
- SCHEINA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, **93**, 10614-10619.
- SCHNEIDER, C., KING, R. M. and PHILIPSON, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787-793.
- SCHNEIDER-MALNOURY, S., GILARDI-HEBENSTREIT, P. and CHARNAY, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819-834.
- SEMENTA, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180-199.
- SEWALL, C. H., BELL, D. A., CLARK, G. C., TRITSCHER, A. M., TULLY, D. B., VANDEN HELVEL, J. and LUCIER, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829-1834.
- SINGH, N., AGRAWAL, S. and RASTOGI, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41-49.
- SMITH, N. R., LI, A., ALDERSLEY, M., HIGH, A. S., MARKHAM, A. F. and ROBINSON, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids Research*, **25**, 3552-3554.
- SOMPATYAC, L., JANZ, S., BURN, T. C., TENEN, D. G. and DANNA, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738-4739.
- ST JOHN, T. P. and DAVIS, R. W., 1979, Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridisation. *Cell*, **16**, 443-452.
- SUN, Y., HEGAMYER, G. and COLBURN, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139-1144.

- ovine aortic smooth muscle
106.
- representational difference
- 1984, Clobazam: species
liver following chronic
- H., 1998, Inhibition and
8, 1203-1253.
- genes associated with high
National Cancer Institute.
- expression by display of 3' end
of Sciences (U.S.A.), 93.
- tion of macrophage gene
ogy Letters, 57, 143-146.
450—an overview. Indian
- : protocol for the selective
thelial cells. *Experimental*
- profiling of non-genotoxic
polymerase chain reaction
etics, 22, 329-333.
- oxisomes and peroxisome
- ALK, W., ANDRESEN, R.,
characterization of colonic
- MHC expression. *Critical*
- on. *Seminars in Oncology*.
- A copy of an RNA species
of DNA. In N. Ford, M.
l. 2nd edition (New York:
- astrula of *Xenopus laevis*.
- W., 1996, Parallel human
genes. *Proceedings of the*
- pressed at growth arrest of
- How to build a vertebrate
10-834.
- isms and pathophysiology.
- VANDEN HEUVEL, J. and
for biomarkers. *Clinical*
- unity: special reference
1-49.
- INSON, P. A., 1997, Rapid
RT-PCR polyacrylamide
- 5, Overcoming limitations
3, 4738-4739.
- e DNA sequences from
ell, 16, 443-452.
- of five messenger RNAs
epidermal cells: one is
Research, 54, 1139-1144.
- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, 23, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler, A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, 192, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, 18, 264-273.
- UTTERLINDEN, A. G., SLACBOOM, P., KNOOK, D. L. and VIJCL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (U.S.A.)*, 86, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, 8, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (U.S.A.)*, 95, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, 270, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AUGUA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, 18, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, 9, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, 14, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (U.S.A.)*, 88, 11505-11509.
- WAWER, C., RUGGEBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, 23, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, 20, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, 223, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (U.S.A.)*, 91, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes: the molecular basis of tumour behaviour. *Journal of Pathology*, 165, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, 187, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, 237, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., NISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, 156, 207-213.
- ZHAO, X. J., NEWSOM, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, 25, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, 21, 709-715.

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*[†], JOHN H. MCCUSKER[‡], AND RONALD W. DAVIS*[§]

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and [†]Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely

important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5' end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1997 by The National Academy of Sciences (X)27-R424/97/948945-3\$2.00/0 PNAS is available online at <http://www.pnas.org>.

[†]Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

[§]To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.

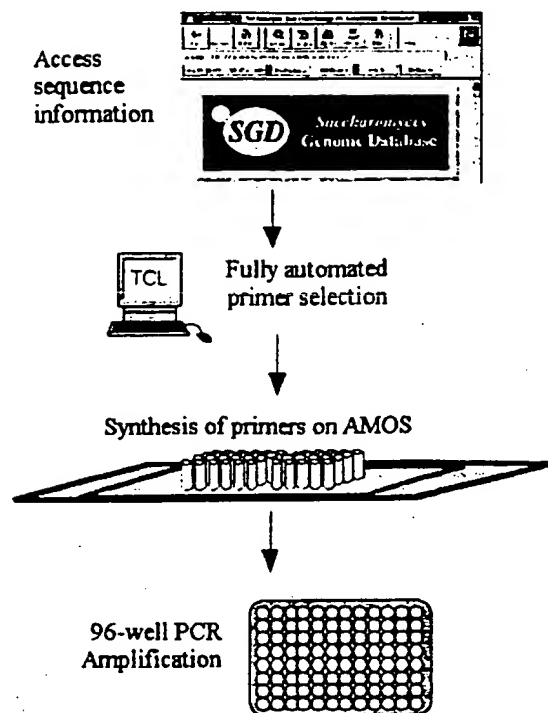


FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose.

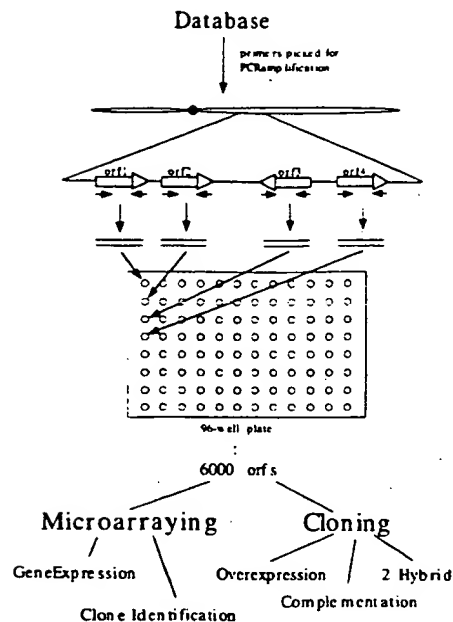


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a “snapshot” of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

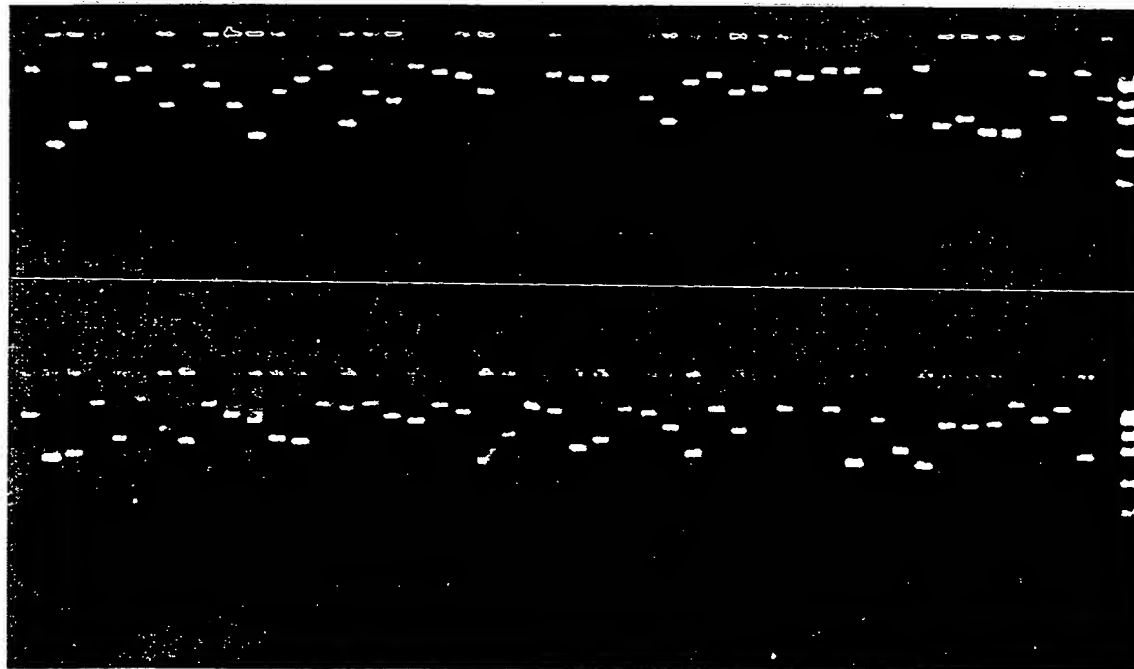


FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct read-out. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

Support was provided by National Institutes of Health Grants R37H60198 and P01H600205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* 269, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* 273, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., & Waterston, R. (1992) *Nature (London)* 356, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* 106, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* 379, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W., & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* 92, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* 57, A266.
10. Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995) *Science* 270, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., & Solas, D. (1991) *Science* 251, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., & Fodor, S. P. (1996) *Science* 274, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M., & Davis, R. W. (1996) *Nat. Genet.* 14, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D., & Brown, P. O. (1996) *Science* 274, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G., & Bandlow, W. (1993) *FEBS Lett.* 316, 41–47.
16. Ramer, S. W., Elledge, S. J., & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* 89, 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E., & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* 340, 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P., & Brown, P. O. (1994) *Nat. Genet.* 4, 11–18.

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,¹ Michael Bittner,² Jeffrey Trent,² J. Carl Barrett,¹ and Cynthia A. Afshari¹

¹Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

²Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22-24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25-27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28-30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only $4n$ cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)⁺ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

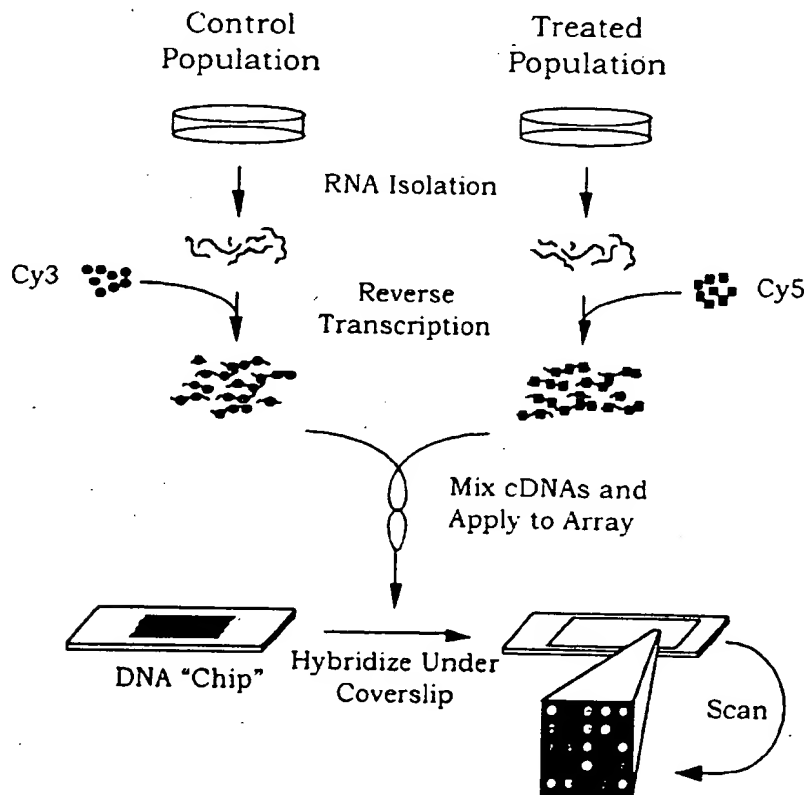


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

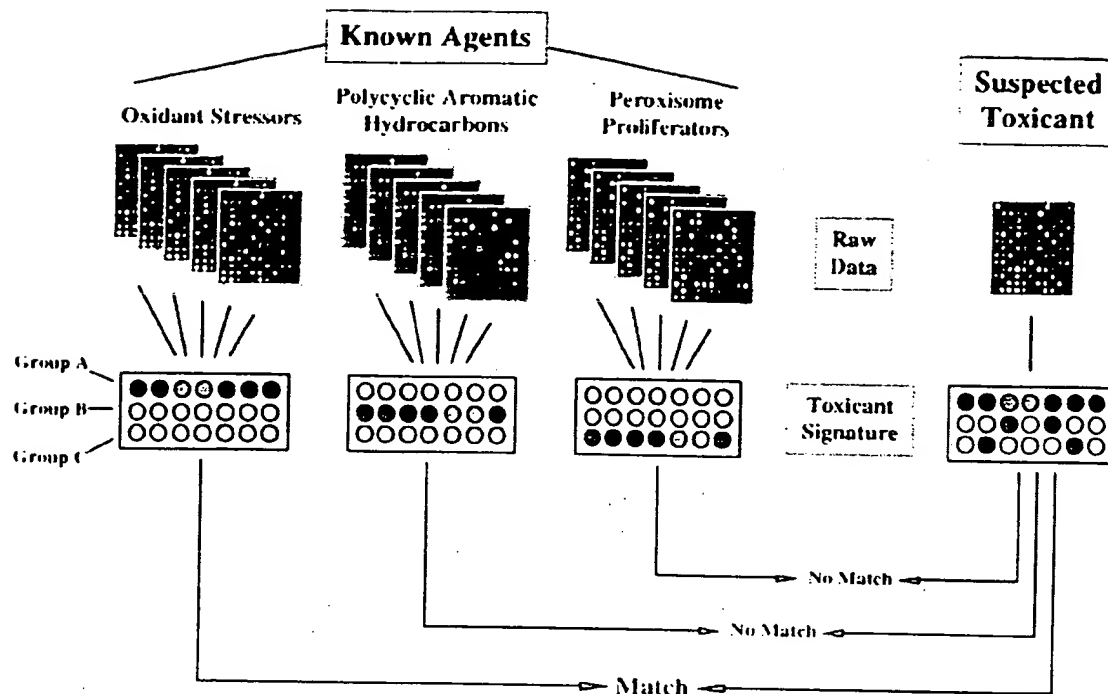


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in *BRCA1* using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (*Gstm1*) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>



Toxicology Letters 112–113 (2000) 467–471

Toxicology Letters

www.elsevier.com/locate/toxlet

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

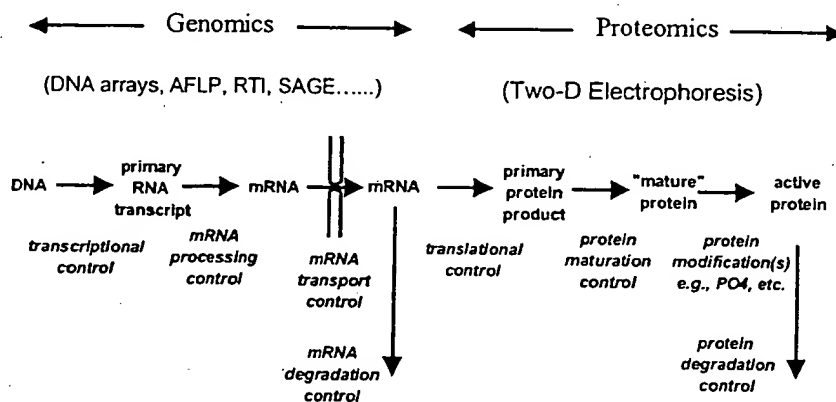


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

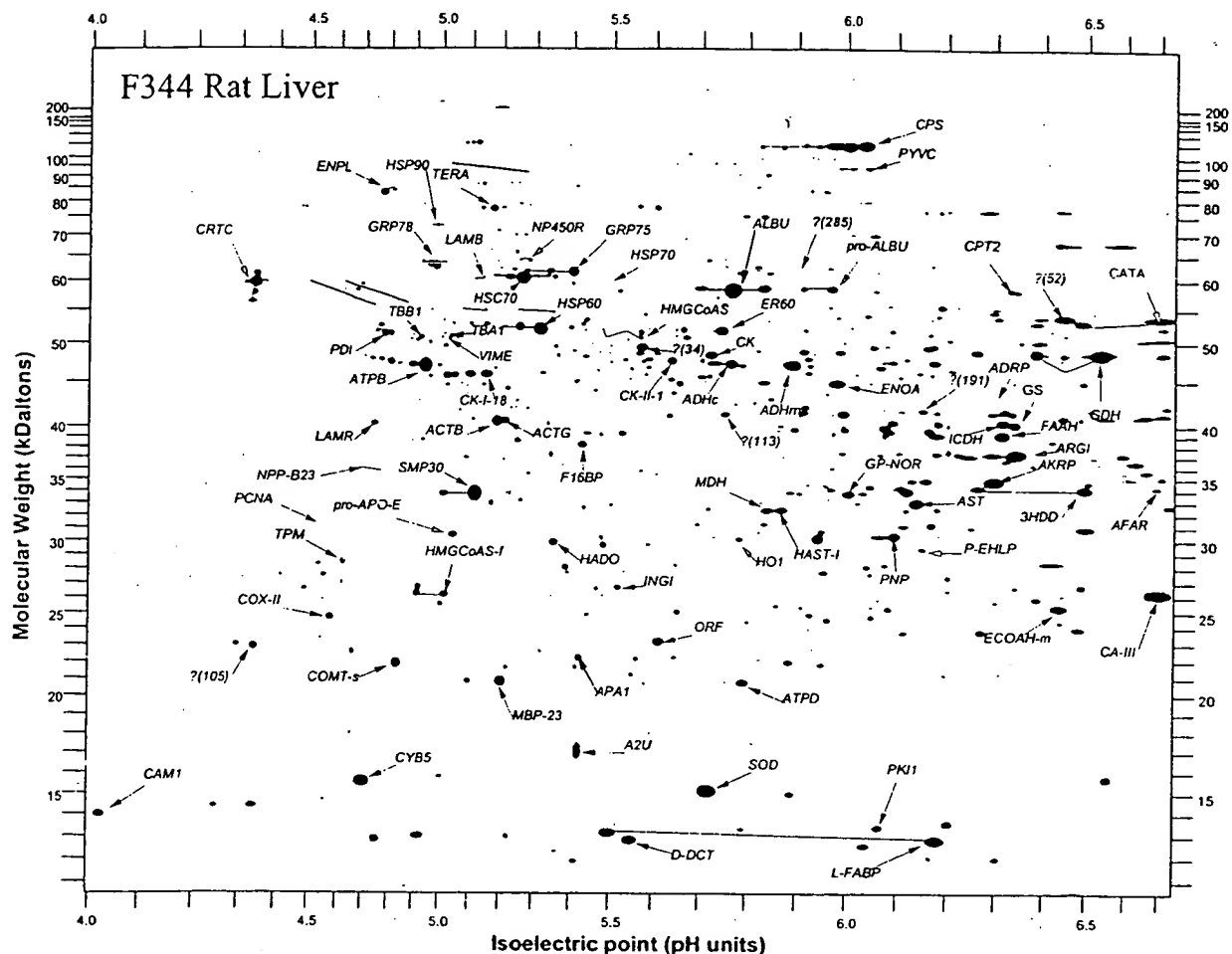


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trials.

References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610-614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355-363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767-773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338-345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157-161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467-470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639-645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raynackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *Biochem. Biophys. Res. Commun.* 218, 777-782.
- Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253-258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543-1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7-8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. **Key words:** DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681-685 (1999). [Online 6 July 1999]
<http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html>

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (1). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the

has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7-8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array Elements

In the context of molecular biology, the word

a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (2). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrays, as described by Brown (3). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of $> 2,500$ spots/cm² may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental

arrays: the first 100 or so spots of a new run tend to be somewhat variable. Factors affecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ³²P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosphorimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (4). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of poly(A)⁺ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After

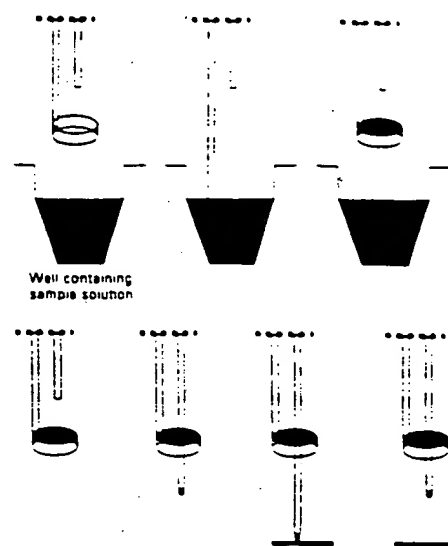


Figure 1. Genetic MicroSystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (14), with permission from Genetic MicroSystems.

normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5' end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden)). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain $> 10^8$ molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (5). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With

Table 1. Advantages and disadvantages of different microarray scanning systems.

	CCD camera system	Nonconfocal laser scanner	Confocal laser scanner
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples	—	May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (13).

elegans knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be

transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.

- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, non-contaminated, sequence-verified, species/cell/tissue/tissue-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a non-related species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings.

Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.

- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

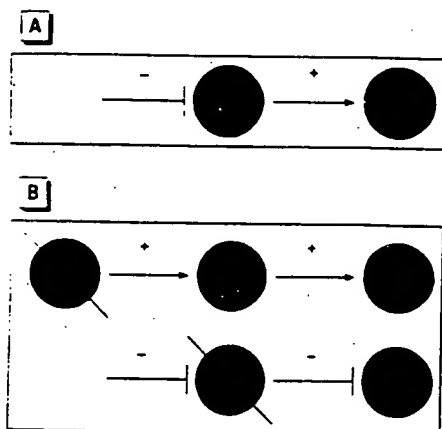


Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2 . (A) A simple, two-component, linear regulatory network operating on gene i_2 , where i_1 is a positive effector of i_2 and j_1 is either a positive or negative effector of i_1 . This network could be deduced by examining the consequence of (B) deleting j_1 on the expression of i_1 and i_2 , where the expression of i_2 would be decreased or increased depending on whether j_1 was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15)

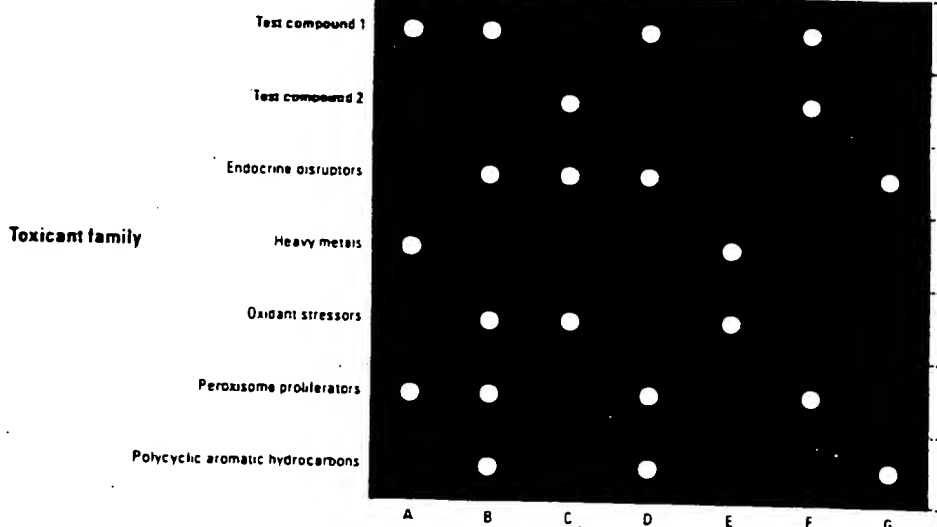


Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results test

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach

is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

REFERENCES AND NOTES

1. The chipping forecast. *Nat Genet* 21(Suppl 1):3-60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Scholar/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: <http://cmgm.stanford.edu/brown> [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* 51:313-324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Maron MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burdard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 4:1293-1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: <http://cmgm.stanford.edu/pbrown/yeastchip.html> [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Alshari CA. Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog* 24(3):153-159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. *Bioessays* 20:555-561 (1998).
10. Zacharewski TR, Timothy R, Zacharewski. Available: www.bch.msu.edu/faculty/zacharewski [cited 22 March 1999].
11. Kramer JA, Krawell SA. RNA in spermatozoa: implications for the alternative haploid genome. *Mol Hum Reprod* 3:473-478 (1997).
12. Stipp D. Gene chip breakthrough. *Fortune*, March 31:56-73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkenen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: <http://www.geneticmicro.com/resources/hml/coldspring.html> [cited 22 March 1999].

SPEAKERS

Cindy Alshari NIHES	Abdel Elkhouloun Research Genetics, Inc.
Linda Birnbaum U.S. EPA	Sue Fenton U.S. EPA
Ron Butow University of Texas Southwestern Medical Center	Norman Hecht University of Pennsylvania
Alex Chenchik Clontech Laboratories, Inc.	Pat Hurban Paradigm Genetics, Inc.
David Duz	Bob Karilock U.S. EPA
	Ernie Kawasaki

Steve Krawetz Wayne State University	Jim Samet U.S. EPA
Nick Mace Genetic Microsystems, Inc.	Sam Ward University of Arizona
Scott Mordacai Allymetrix, Inc.	Jeff Welch U.S. EPA
Kevin Morgan Glaxo Wellcome, Inc.	Reen Wu University of California at Davis
Elaine Poplin Research Genetics, Inc.	Tim Zacharewski Michigan State University
Don Rose	

Subject: RE: [Fwd: Toxicology Chip]

Date: Mon, 3 Jul 2000 08:09:45 -0400

From: "Afshari,Cynthia" <afshari@niehs.nih.gov>

To: "'Diana Hamlet-Cox'" <dianahc@incyte.com>

You can see the list of clones that we have on our 10K chip at
<http://marvel.niehs.nih.gov/maps/guest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80-) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80- genes across our experiments.

Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.

I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox

> Sent: Monday, June 26, 2000 8:52 PM

> To: afshari@niehs.nih.gov

> Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not
> the right person to contact.

> Can you help me in this matter? I don't need to know the sequences,
> necessarily, but I would like very much to know what types of sequences
> are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----

> Subject: Toxicology Chip

> Date: Mon, 19 Jun 2000 18:31:48 -0700

> From: Diana Hamlet-Cox <dianahc@incyte.com>

> Organization: Incyte Pharmaceuticals

> To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as
> pharmacotoxicology markers, and found the Press Release dated February
> 29, 2000 regarding the work of the NIEHS in this area. I would like to
> know if there is a resource I can access (or you could provide?) that
> would give me a list of the 12,000 genes that are on your Human ToxChip
> Microarray. In particular, I am interested in the criteria used to
> select sequences for the ToxChip, including any control sequences
> included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.

> Incyte Genomics, Inc.

> --

> =====

wd Toxicology Cntrl

> This email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.

> =====

>

>

- These data were digitally recorded by stand-alone units [borrowed from Incorporated Research Institutions for Seismology (IRIS), Washington, DC] arranged in arrays 25 km long with receiver intervals of 100 to 300 m.
20. Gravity was measured by a Worden (Texas Instruments, Houston, TX) gravimeter and tied to the McMurdo gravity base station. The observed gravity anomaly was found to be within 0.5 mgal of that from a previous regional survey (2). Relative elevations along the profile were measured by pairs of barometers.
 21. Holes were drilled by melting of the ice with hot water pumped under pressure. For the multichannel reflection work, 7.5-kg charges were spaced every 200 m along the profile. Two charges 1.6 km apart were detonated for each streamer location. This pattern resulted in an effective 120-channel receiving array 3 km long. The sources for the wide-angle reflection and refraction data were several explosions of 100 to 400 kg.
 22. L. R. Bartek, P. R. Vail, J. B. Anderson, P. A. Emmet, S. Wu, *J. Geophys. Res.* **96**, 6753 (1991).
 23. K. Hinz and M. Block, in *Proceedings of the 11th World Petroleum Congress* (Wiley, New York, 1984), pp. 279-291.
 24. Velocities directly beneath the sea floor were less than that of solid ice (3.8 km s^{-1}) and therefore produced no refracted arrivals. For this portion of the refraction model (beneath the sea floor under the ice shelf), velocities were obtained from the stacking velocities of the reflection data.
 25. A. K. Cooper and F. J. Davey, *The Antarctic Continental Margin: Geology and Geophysics of the Western Ross Sea* (Circum-Pacific Council for Energy and Mineral Resources (CPCEMR), Houston, 1987).
 26. D. R. H. O'Connell and T. M. Stepp, in *Geol. Jahrb.*, in press.
 27. L. D. Hale and G. A. Thompson, *J. Geophys. Res.* **87**, 4625 (1982).
 28. R. Meissner and N. J. Kusznir, *Ann. Geophys. Ser. B* **5**, 365 (1987).
 29. B. S. Gibson and A. R. Levander, *Geophys. Res. Lett.* **15**, 617 (1988).
 30. W. A. Heiskanen and F. A. Vening Meinesz, *The Earth and its Gravity Field* (McGraw-Hill, New York, 1958).
 31. J. Braun and C. Beaumont, *Geology* **17**, 760 (1989).
 32. R. S. Stein, G. C. P. King, J. B. Rundle, *J. Geophys. Res.* **93**, 13319 (1988).
 33. M. H. P. Bott and T. A. Stern, *Tectonophysics* **201**, 341 (1992).
 34. R. I. Kalamirides, J. H. Berg, R. A. Hank, *Science* **237**, 1192 (1987).
 35. E. B. Goodwin and J. McCarthy, *J. Geophys. Res.* **95**, 20097 (1990).
 36. W. R. Buck, *Earth Planet. Sci. Lett.* **77**, 362 (1986).
 37. P. J. Barrett, M. J. Hambrey, D. M. Harwood, A. R. Pyne, P.-N. Webb, *DSIR Bull.* **245**, 241 (1989).
 38. C. M. Clapperton and D. E. Sugden, *Quat. Sci. Rev.* **9**, 253 (1990).
 39. A. C. Johnston, *Nature* **330**, 467 (1987).
 40. G. Zandt and G. Owens, *Bull. Seismol. Soc. Am.* **70**, 1501 (1980).
 41. P. Wellman and R. J. Tingey, *Nature* **291**, 142 (1981).
 42. The stacked profile was enhanced by the combination of four adjacent traces and the application of a two-dimensional median filter and was not migrated.
 43. The BSR here shares similar characteristics with BSRs identified and drilled in low-latitude margins: It simulates the sea floor, cuts across stratigraphic units, exhibits a polarity reversal of the wavelet compared to the sea floor, and is overlain by an acoustically opaque layer. Bottom-simulating reflectors are caused by a thin layer of free gas at the base of a gas hydrate-cemented layer. Methane gas has been encountered in drill holes and cores around Antarctica, but despite the anomalously large depth of the Antarctic shelves, BSRs have rarely been documented [K. A. Kvenvolden, M. Golan-Bac, J. B. Rapp, in *The Antarctic Continental Margin: Geology and Geophysics of Offshore Wilkes*

- Land*, S. L. Eitrem and M. A. Hampton, Eds. (CPCEMR, Houston, 1987), pp. 205-213].
44. We modeled the refraction data using both forward- and inverse-ray tracing programs [J. H. Luetgert, *U.S. Geol. Surv. Open-File Rep.* **88-238** (1988), p. 52; W. J. Lutter, R. L. Nowack, L. W. Braile, *J. Geophys. Res.* **95**, 4621 (1990)]. Values for ice thickness and absolute elevation from previous regional surveys of the Ross Ice Shelf [D. G. Albert and C. R. Bentley, in *The Ross Ice Shelf: Glaciology and Geophysics*, C. R. Bentley and D. E. Hayes, Eds. (American Geophysical Union, Washington, DC, 1990), pp. 87-108] were used for modeling of the refraction and gravity data.
 45. We thank R. Busby, B. Harris, S. Heaphy, T. Hefford,

C. Hobbs, N. Lord, D. Lousley, and B. Straite for their invaluable help in the field; Polar Ice Coring Office, (University of Alaska, Fairbanks) for drilling the shot holes; A. Melhuish, M. Lee, W. Agena, W. Poag, W. Dillon, and J. Zwinakis for helping with processing, interpretation, and drafting; P. Barrett, C. Bentley, F. Davey, D. Hutchinson, K. Klitgord, G. Thompson, and one anonymous individual for helpful reviews; Norsk-Hydro, Y. Kristoffersen, and E. Rygg for lending the snow streamer; Grant-Norpac (Houston, TX) for lending computer boards; and IRIS for sending a technician and lending stand-alone recording units and peripherals. Funded by National Science Foundation grant DPP89-17634 and the New Zealand Science Foundation.

RESEARCH ARTICLES

Three-Dimensional Structure of Myosin Subfragment-1: A Molecular Motor

Ivan Rayment,* Wojciech R. Rypniewski,† Karen Schmidt-Bäse,‡ Robert Smith, Diana R. Tomchick,§ Matthew M. Benning, Donald A. Winkelmann, Gary Wesenberg, Hazel M. Holden

Directed movement is a characteristic of many living organisms and occurs as a result of the transformation of chemical energy into mechanical energy. Myosin is one of three families of molecular motors that are responsible for cellular motility. The three-dimensional structure of the head portion of myosin, or subfragment-1, which contains both the actin and nucleotide binding sites, is described. This structure of a molecular motor was determined by single crystal x-ray diffraction. The data provide a structural framework for understanding the molecular basis of motility.

Motility is one of the characteristic features of many living organisms and involves the transduction of chemical into mechanical energy. Only a limited number of strategies have evolved to accomplish this task. At present, three major classes of molecular motors have been identified, myosin, dynein, and kinesin, and all are important in cellular movement (1). Of these three proteins, the most abundant is myosin, which plays both a structural and an enzymatic role in both muscle contraction and intracellular motility.

I. Rayment, W. R. Rypniewski, K. Schmidt-Bäse, R. Smith, D. R. Tomchick, M. M. Benning, G. Wesenberg, and H. M. Holden are in the Department of Biochemistry and Institute for Enzyme Research, University of Wisconsin, 1710 University Avenue, Madison, WI 53705. D. A. Winkelmann is in the Department of Pathology, Robert Wood Johnson Medical School, Piscataway, NJ 08854.

*To whom all correspondence should be addressed.
 †Present address: European Molecular Biology Laboratory, Notkestrasse 85, 2000 Hamburg 52, Germany.
 ‡Present address: Max Planck Institute for Biochemistry, 8033 Martinsried, Germany.
 §Present address: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907.

The role of myosin in movement has been most clearly defined from the study of cross-striated skeletal muscle, which shows a high degree of structural organization. In striated muscle the basic contractile unit is the sarcomere, which consists of overlapping arrays of thick and thin filaments. During contraction, these filaments, which are composed primarily of myosin and actin, respectively, slide past one another, thereby shortening the length of the sarcomere (2). Electron micrographs of muscle in rigor have revealed connections between the filaments in the overlap region, the so-called crossbridges. These crossbridges are formed by the globular regions of the myosin molecule and are responsible for force generation in the contractile process through the hydrolysis of adenosine triphosphate (ATP).

Myosin, which has a molecular size of about 520 kilodaltons, consists of two 220-kD heavy chains and two pairs of light chains that vary in molecular size depending on the source but are usually between 15 and 22 kD (3, 4). The molecule is highly

asymmetric, consisting of two globular heads attached to a long tail. Each heavy chain forms the bulk of one head and intertwines with its neighbor to form the tail. Limited proteolytic digestion has shown that the myosin head, or subfragment-1 (S1), contains an ATP, actin, and two light chain binding sites and that the myosin rod, which is formed by a coiled coil of two α helices, accounts for the self-association of myosin at low ionic strength and the formation of the thick filament backbone (3). Spudich and co-workers have demonstrated that the globular head portions of myosin are sufficient to generate movement of actin in an *in vitro* motility assay (5).

Each globular head, derived from limited proteolysis, consists of a heavy chain fragment having a molecular size of 95 kD and two light chains yielding a combined molecular size of ~130 kD (6). The two light chains differ in their structure and properties and are known by a variety of names. In this article they are referred to as the regulatory and essential light chains. Neither type is required for the adenosine triphosphatase (ATPase) activity of the head (7). In some species, however, these chains regulate or modulate the ATPase activity of myosin in the presence of actin (8, 9). Amino acid sequence analyses reveal that both light chains share considerable sequence similarity with calmodulin and troponin C although most of the divalent cation binding sites have been lost during evolution (10).

During the last 40 years, enormous effort has been expended toward understanding the structure and function of the myosin head (11). Measurements from electron micrographs have suggested that the myosin head is pear-shaped, about 190 Å long and 50 Å wide at its thickest point (12). Molecular dimensions subsequently derived from studies of fixed thin sections cut from crystals of myosin S1 were consistent with these observations (13).

Although much biochemical and physical information has accumulated for myosin since 1950, structural knowledge of this protein or any other molecular motor has been lacking. We now describe the tertiary structure of the myosin head and suggest how this protein may serve to transduce energy from the hydrolysis of ATP into directed movement. We present the three-dimensional structure of myosin S1 at a nominal resolution of 2.8 Å and refinement R factor of 22.3 percent for all x-ray data recorded in that range.

Crystallization of myosin subfragment-1. Myosin is an abundant protein that can be easily prepared in gram quantities. Likewise, the myosin head, which is readily cleaved from the rest of the molecule by mild prote-

olysis, can be prepared in large quantities. This soluble subfragment has been known for approximately 30 years and has resisted crystallization despite numerous attempts. In view of its central importance for understanding the molecular basis of muscle contraction, we undertook an alternative approach to the usual ways of obtaining x-ray quality crystals. The protein was first subjected to mild chemical modification of the lysine residues by reductive methylation. This chemical modification has long been used as a gentle way to introduce a radioactive label into a protein (14).

Considerable effort was expended to determine the optimal procedure for modifying the protein since it was recognized that complete, homogeneous modification of the molecule was essential for obtaining high-quality crystals (Table 1). Many of the experiments necessary to derive the optimal protocol for methylation were performed in a parallel study on hen egg white lysozyme (15). In that study the three-dimensional structure of the modified protein was determined and refined to 1.8 Å resolution and shown to be essentially identical to that of the native protein except for the modified lysine residues. Modification of the lysine residues in

Table 1. Amino acid analysis of modified and native myosin S1 (60). Prior to modification, the protein, at 5 mg/ml, was dialyzed against 200 mM potassium phosphate, pH 7.5, 1 mM $MgCl_2$. The protein was reductively methylated at 4°C by the sequential addition of 1 M dimethylamine borane complex dissolved in water (20 μ l per milliliter of protein) and 1 M formaldehyde (40 μ l per milliliter of protein) with rapid stirring. This process was repeated after 2 hours; a further portion (10 μ l/ml) of dimethylamine borane complex was added after 2 hours and the reaction mixture was kept overnight at 4°C in the dark. The reaction was quenched by the addition of 3.8 M ammonium sulfate to a final concentration of 1 M and then dialyzed for 48 hours against 2.5 M ammonium sulfate, 50 mM potassium phosphate at pH 6.7 to precipitate the protein (15, 49). All except three to four of the lysine residues were modified. Discrepancy between the total number of lysine residues in the native and modified protein may have arisen from a calibration error in the dimethyllysine standard. The analyses for histidine, methionine, and arginine are shown as controls.

Amino acid	Residues (no.)		
	Theoretical	Native	Modified
Lysine	103	96.2	4.2
Me ₁ -Lys	0	0	0
Me ₂ -Lys	0	0.6	96.7
Me ₃ -Lys	3	3.6	3.4
Total lysine	106	100.4	104.3
Histidine	24	23.7	23.2
Methionine	39	39.4	38.9
Arginine	46	47.5	47.2

lysozyme dramatically changed its crystallization properties. The kinetic and structural effects of this treatment on myosin S1 are discussed below.

Myosin isolated from chicken pectoralis muscle consists of a mixed population of two isozymes caused by the existence of two species of the essential light chain (16). These light chains are referred to as A1 (21 kD) and A2 (16 kD). Amino acid sequence studies of the light chains have demonstrated that A1 and A2 are identical over their 142 residues at the COOH-terminus. The size difference is caused by an additional 41 amino acids present at the NH₂-terminus of A1. These isozymes arise by alternative transcription and two modes of splicing from a single gene (17).

The crystals used in our study contained both isoforms of the essential light chain. Myosin S1 was prepared by digestion with papain in the presence of $MgCl_2$ because the fragment produced under these conditions contained both the regulatory and essential light chains. The major drawback of papain as a proteolytic enzyme, however, was its lack of specificity. Apart from cleaving the heavy chain at the head-rod junction, additional proteolytic breaks were introduced into both the regulatory and A1 essential light chains. Also, there was partial phosphorylation of the regulatory light chain by endogenous myosin light chain kinase. The myosin S1 was prepared by an improved purification protocol that removed the heterogeneity arising from both proteolysis of the light chains and phosphorylation of the regulatory light chain (18).

Crystals were grown by batch methods from 1.35 M ammonium sulfate, 500 mM potassium chloride, and 50 mM potassium phosphate (pH 6.7) in the presence of 5 mM dithiothreitol and 0.5 mM sodium azide at a final protein concentration of 8 to 12 mg/ml. Crystallization was initiated by microseeding, and the crystals grew as thick rods to a length of 1 to 2 mm and a width and thickness of 0.4 and 0.3 mm, respectively, over a period of 2 to 3 months at 4°C. They belonged to the space group C222₁ with unit cell dimensions of $a = 98.4$, $b = 124.2$, $c = 274.9$ Å, and one molecule in the asymmetric unit. These crystals were different from those originally reported (19) and arose from improvements in both the chemical modification procedure and the protein homogeneity.

Structure determination. The x-ray data were collected in two stages (20). First, x-ray data sets to 4.5 Å resolution for the native and heavy atom-containing crystals were recorded by an area detector with the goal of determining the positions of the metal binding sites. These data were then extended to 2.8 Å resolution with

synchrotron radiation at Stanford University (SSRL) and Cornell University (CHESS). We recognized early that x-ray data collection and determination of the protein phases by multiple isomorphous replacement would be difficult unless care was taken to minimize the systematic errors introduced by differences between the successive protein preparations. Consequently, for each stage in the heavy atom derivative data collection, a corresponding native data set was recorded from the same protein preparation. For each purification trial, approximately 700 mg of myosin S1 was prepared and set up for crystallization. Many attempts were made before a single preparation yielded sufficient crystals for x-ray data collection.

The structure was determined by a combination of multiple isomorphous replacement and solvent flattening. The first derivative solved was obtained from crystals soaked in trimethyllead acetate and proved to be highly isomorphous with only four binding sites (21). It was used to determine the positions of the other heavy atom binding sites by difference Fourier techniques (Table 2). The positions and occupancies of the heavy atom sites were refined according to the origin-removed Patterson-function correlation method by the program HEAVY (22). The overall figures of merit for the area detector, CHESS, and SSRL synchrotron data were 0.47, 0.58, and 0.42, respectively.

The higher resolution x-ray data collected at SSRL were placed on the same scale as the area detector data and included as a block from 4.5 to 2.8 Å. Efforts to merge the overlapping data between the area detector and synchrotron data were unsatisfactory. However, the phase information from all three sources was combined throughout the entire resolution range via the phase probability coefficients (23).

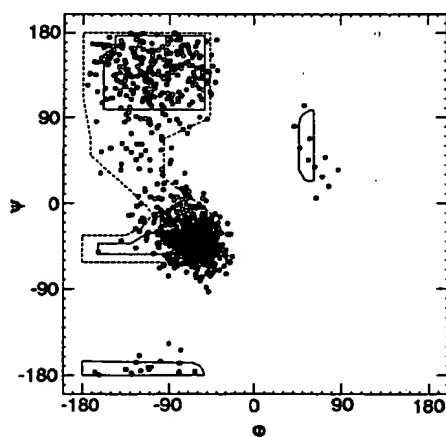


Fig. 1. Ramachandran plot of the main chain dihedral angles of all non-glycyl residues in the model presented.

These protein phases were improved by solvent flattening (24). The positions and occupancies of heavy atom binding sites were further refined against these modified phases (25). This gave an improved electron density map into which approximately 550 alanine residues were built with the program FRODO (26). The map showed good connectivity and many well-defined side chains.

Once several long segments were connected, the positions of these alanine residues were matched to the known amino acid sequence (27, 28). At this stage phase information from the partial model was combined with the heavy atom derivative phases by the program SIGMAA (29). The structure was refined concurrently with the model building process by the program package TNT (30). Once the model building was near completion, a cycle of refinement with X-PLOR (31) was performed to improve the conformations of the side chains. The strategy of alternate model

building and refining proved successful and constantly improved the estimation of the protein phases. Toward the end of the analysis there were clear segments of electron density corresponding to portions of the light chains that were completely missing in the original maps phased with heavy atom derivatives alone.

At present, 1072 residues (of a total of 1157) have been built into the electron density map. The model was refined to an *R* factor of 22.3 percent for all measured x-ray data between 30 to 2.8 Å with root-mean-square deviations from ideal geometry of 0.018 Å for bond lengths, 2.5° for bond angles, and 0.013 Å for groups of atoms expected to be coplanar. No solvent molecules have yet been built into the electron density (Figs. 1 and 2).

Structure description. In a space-filling representation of all atoms in the myosin S1 model (Fig. 3), the green, red, and blue segments represent parts of the heavy chain and the yellow and magenta stretches cor-

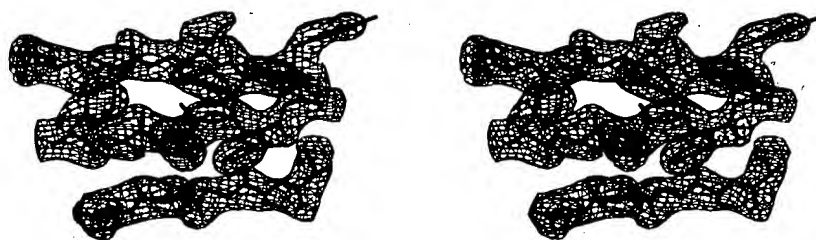


Fig. 2. A stereo view of a representative section of electron density located in the seven-stranded β sheet motif of the heavy chain calculated with SIGMAA coefficients (29). The phases and weights used to calculate the electron density were obtained by combining the information from the heavy atom phases and those derived from the atomic model.

Table 2. Heavy atom derivatives used in the structure determination and their data collection statistics.

Derivative	Conditions*		Method	R_{sym}^{\dagger}	Reflections (no.)	Resolution (Å)	$R_{\text{scale}}^{\ddagger}$	Sites (no.)	Phasing power §
	Concentration (mM)	Time (days)							
Trimethyllead acetate	20	21	Area detector	6.7	13,394	3.5	24.5	4	1.01
KAu(CN) ₂	1	5	Area detector	5.8	18,082	3.5	18.2	6	1.14
K ₂ OsO ₄ /pyridine	2–20	2	Area detector	6.7	11,804	4.0	24.5	4	1.02
K ₃ UO ₂ F ₅	3	4	Area detector	6.7	11,688	4.0	19.1	7	1.06
Trimethyllead acetate	20	21	CHESS	11.5	34,498	2.8	28.2	4	1.19
KAu(CN) ₂	1	5	CHESS	9.7	36,339	2.8	17.0	8	1.23
K ₂ OsO ₄ /pyridine	2–20	2	CHESS	10.0	32,554	2.8	32.2	8	0.99
Cis-Pt(NH ₃) ₂ Cl ₂	2	3	CHESS	10.3	36,108	2.8	21.0	12	1.12
K ₃ UO ₂ F ₅	3	4	CHESS	10.9	36,419	2.8	22.7	9	0.98
Trimethyllead acetate	15	21	SSRL	13.0	31,667	2.8	27.1	4	1.11
K ₃ UO ₂ F ₅	2	3	SSRL	11.8	33,043	2.8	22.2	6	0.88

*The heavy atom derivatives were prepared at 4°C by first slowly transferring the crystals to a synthetic mother liquor composed of 1.5 M ammonium sulfate, 500 mM KCl buffered with 20 mM Pipes at pH 6.7. $\dagger R_{\text{sym}} = \sum \sum (|I_{hi}| - |I_{hi}|) / \sum \sum I_{hi} \times 100$, where I_{hi} and I_h are the intensities of the individual and mean structure factors. $\ddagger R_{\text{scale}} = \sum (|F_{hi}| - |F_{hi}|) / \sum F_{hi} \times 100$, where F_{hi} and F_h are the heavy atom and native structure factors. \S The phasing power is defined as the mean value of the heavy atom structure factor divided by the residual lack-of-closure error.

respond to the essential and regulatory light chains, respectively. As can be seen, the myosin head is highly asymmetric with a length of 165 Å, a width of 65 Å, and a thickness of approximately 40 Å.

Previous knowledge of the organization of the heavy chain in the myosin head was derived from proteolytic studies. Limited tryptic digestion of vertebrate skeletal S1 indicated that the head contained three major regions: a 25-kD NH₂-terminal nucleotide binding region (32), a central 50-kD segment, and a 20-kD COOH-terminal segment; the last two were shown to bind to

actin (33, 34). These proteolytic segments are displayed in green, red, and blue, respectively (Fig. 3); the light chains abut one another and are wrapped around a single α helix of the heavy chain but do not overlap to any significant extent.

The secondary structure of the myosin head is dominated by α helices with approximately 48 percent of the amino acid residues in this conformation (Figs. 4 and 5). One key structural feature is the long (approximately 85 Å) α helix which extends from the thick part of the head down to the COOH-terminus of the heavy chain.

This α helix constitutes the light chain binding region of the heavy chain. There is a bend, delineated by amino acid residues Trp⁸²⁹, Pro⁸³⁰, Trp⁸³¹, and Met⁸³², which connects this long α helix to a short COOH-terminal α helix of the 95-kD heavy chain fragment. A brief description of the three polypeptide chains constituting the myosin head is given below.

The regulatory light chain is located at the end of the molecule distal from the nucleotide binding site (Figs. 4 and 5). It consists of two domains and shares considerable structural homology with calmodulin and troponin C except that the long connecting helix observed in calmodulin and troponin C is distorted (35, 36). A comparison of the regulatory light chain with calmodulin is shown in Fig. 6A where the eight helices that comprise the two domains have been labeled A through H. The regulatory light chain is arranged such that its NH₂-terminal domain wraps around the COOH-terminus of the heavy chain between amino acid residues Asn⁸²⁵ and Leu⁸⁴² whereas its COOH-terminal domain interacts with the heavy chain in the region defined by amino acid residues Glu⁸⁰⁸ to Val⁸²⁶. The interaction of the NH₂-terminal domain with the heavy chain is stabilized by a cluster of hydrophobic residues including nine phenylalanines, two trypt-

Fig. 3. A space-filling representation of all of the atoms in the current model of myosin S1. The model is oriented such that the actin binding surface is located at the lower right-hand corner. The 25-, 50-, and 20-kD segments of the heavy chain are colored in green, red, and blue, respectively, whereas the essential and regulatory light chains are shown in yellow and magenta, respectively. In this orientation the prominent horizontal cleft that divides the central 50-kD segment of the heavy chain into two domains (upper and lower defined by this orientation) is clearly visible. This figure was prepared with the molecular graphics program MIDAS (61).

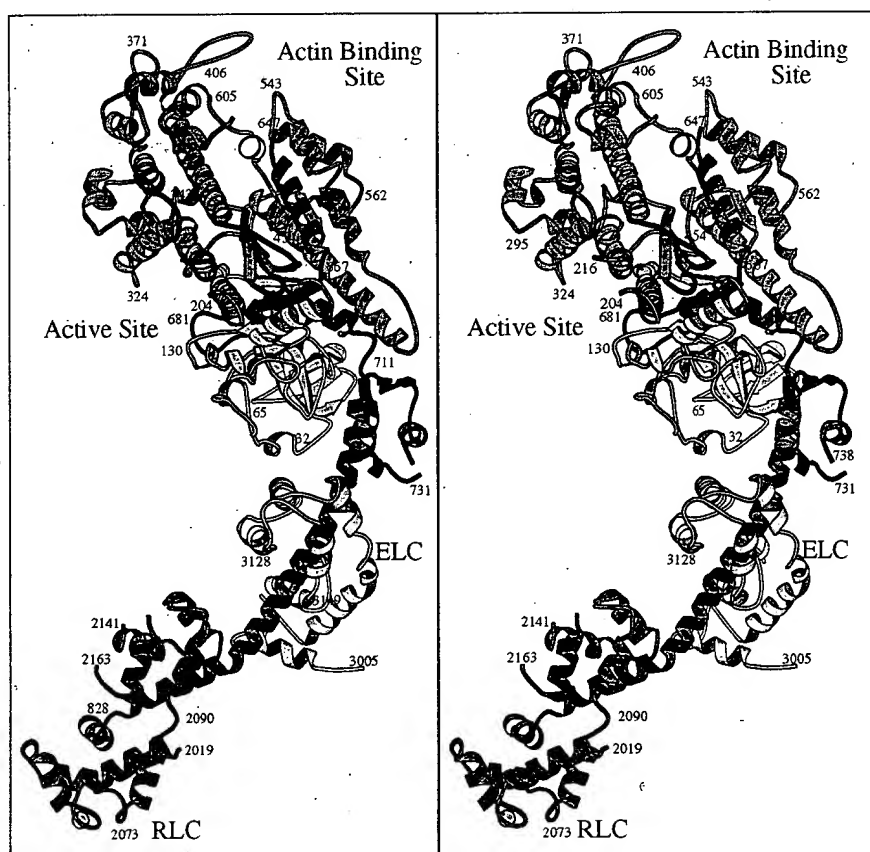
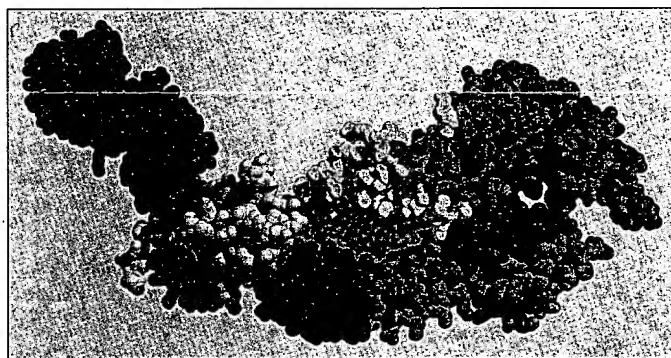


Fig. 4. A ribbon representation of the entire model for myosin S1. In this and all successive figures, 2000 and 3000 have been added to the residue numbers of the regulatory and essential light chains, respectively, to distinguish these from the heavy chain. Heavy chain residues Asp⁴ to Glu²⁰⁴, Gly²¹⁶ to Tyr⁶²⁶, and Gln⁶⁴⁷ to Lys⁸⁴³ are colored in green, red, and blue, respectively. These segments are separated by disordered loops for which no density is evident in the current map. There are two additional segments in the heavy chain for which the density is weak or disordered. These include residues Lys⁵⁷² to Lys⁵⁷⁴ and Ile⁷³² to Phe⁷³⁷. The A2 isozyme of the essential light chain, shown in yellow, theoretically contains 149 amino acid residues. In the model it extends from residue Asp⁵ to Val¹⁴⁹ and contains one ill-defined region that includes residues Leu⁵⁰ to Ala⁶⁰. The regulatory light chain, which is colored in magenta, theoretically consists of 166 amino acid residues. In the current model it extends from residue Phe¹⁹ to Lys¹⁶³ but is disordered between residues Pro¹⁴² and Asn¹⁴⁷. In this figure the molecule is oriented perpendicular to its long axis and rotated to view along the active site pocket. A sulfate ion, shown here in a space-filling representation, is located at the base of the pocket. The actin binding surface has been defined as indicated on the figure by the location of the 50- to 20-kD junction (residues Tyr⁸²⁶ and Gln⁶⁴⁷) and by its interaction with actin (46). Figures 4 to 7 were prepared with the molecular graphics program MOLSCRIPT (62).

tophans, and four methionines. Five of these residues are contributed by the heavy chain. Superposition of the NH₂-terminal

domains of the regulatory light chain and calmodulin reveals an rms difference in the positions of 59 equivalent residues of only

1.3 Å. By contrast the COOH-terminal domain is less similar to the structure observed in calmodulin and is due to a difference in the positions of the F and G helices in the regulatory light chain that have moved to accommodate the heavy chain. In addition, the COOH-terminal domain as a whole has rotated, relative to calmodulin, about the midpoint between the two domains in order to form a tight complex with the heavy chain.

The divalent cation binding site is located in the first helix-loop-helix motif observed in the amino acid sequence and, as indicated above, has a conformation similar to that observed in calmodulin. A divalent cation is clearly evident in the electron density and is most likely Mg²⁺ in that this was a minor constituent of the crystallization buffer. In our model, no electron density was observed for the first 18 amino acid residues in the regulatory light chain. This includes Ser¹³ and is, by sequence homology to rabbit myosin, the site of phosphorylation by myosin light chain kinase (37). Presumably this portion of the polypeptide chain is flexible in S1 and perhaps only plays a functional role when the head is attached to the remainder of the molecule. The observed NH₂- and COOH-terminal residues of the regulatory light chain lie close to the interface between the two domains.

The essential light chain interacts with the long α helix of the heavy chain through amino acid residues Leu⁷⁸³ to Met⁸⁰⁶ (Fig. 6C). Likewise, it wraps around the heavy chain α helix but in a manner different from that observed for the regulatory light chain. Its arrangement resembles that for the interaction of calmodulin with a target peptide from myosin light chain kinase (38). It differs in that the second and third helices in the NH₂-terminal domain abut the heavy chain with their external surfaces, whereas the corresponding secondary structural elements in calmodulin enclose the respective target peptide. The electron density for this part of the molecule is the least well ordered of the entire map. Indeed, very little of the essential light chain was visible in the original electron density map and only appeared after the phase information from the rest of the molecule was included. This could be due to either lack of isomorphism in the heavy atom derivative phases or conformational flexibility of the polypeptide chain. It is difficult to distinguish between these two possibilities because the crystals contain both classes of essential light chain isoforms. As with the regulatory light chain, the NH₂- and COOH-terminal residues lie close to the interface between the two domains.

The heavy chain constitutes the entire thick portion of the myosin head and con-

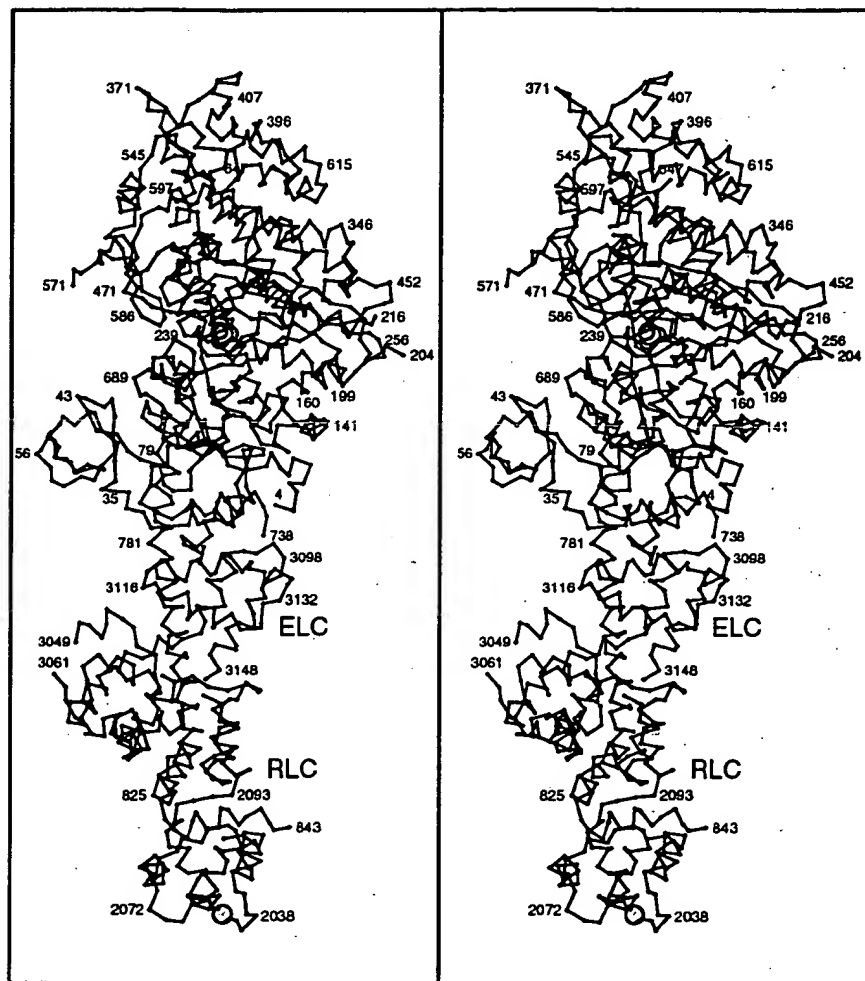


Fig. 5. A stereo α carbon plot of the entire myosin head in which the view has been rotated 90° with respect to Fig. 4. In this view, the active site pocket is seen as a wide depression. Selected residues have been labeled to allow the path of the chain to be followed and to identify the start and end of the secondary structural elements.

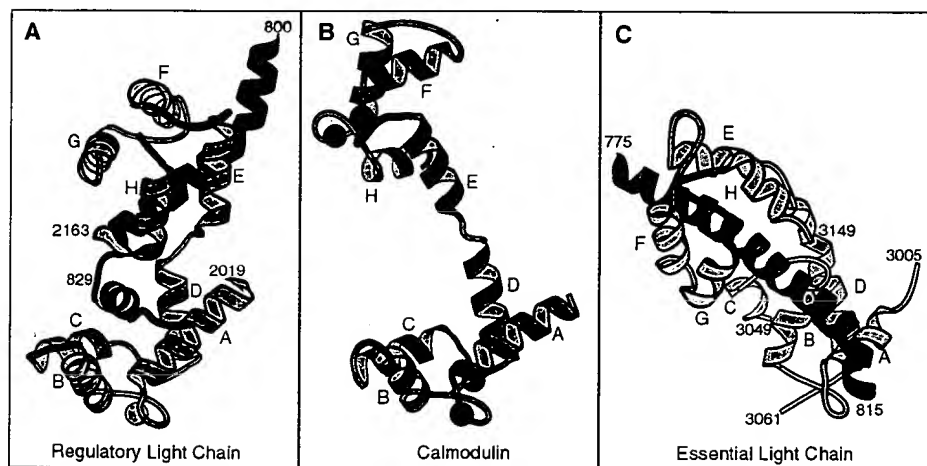
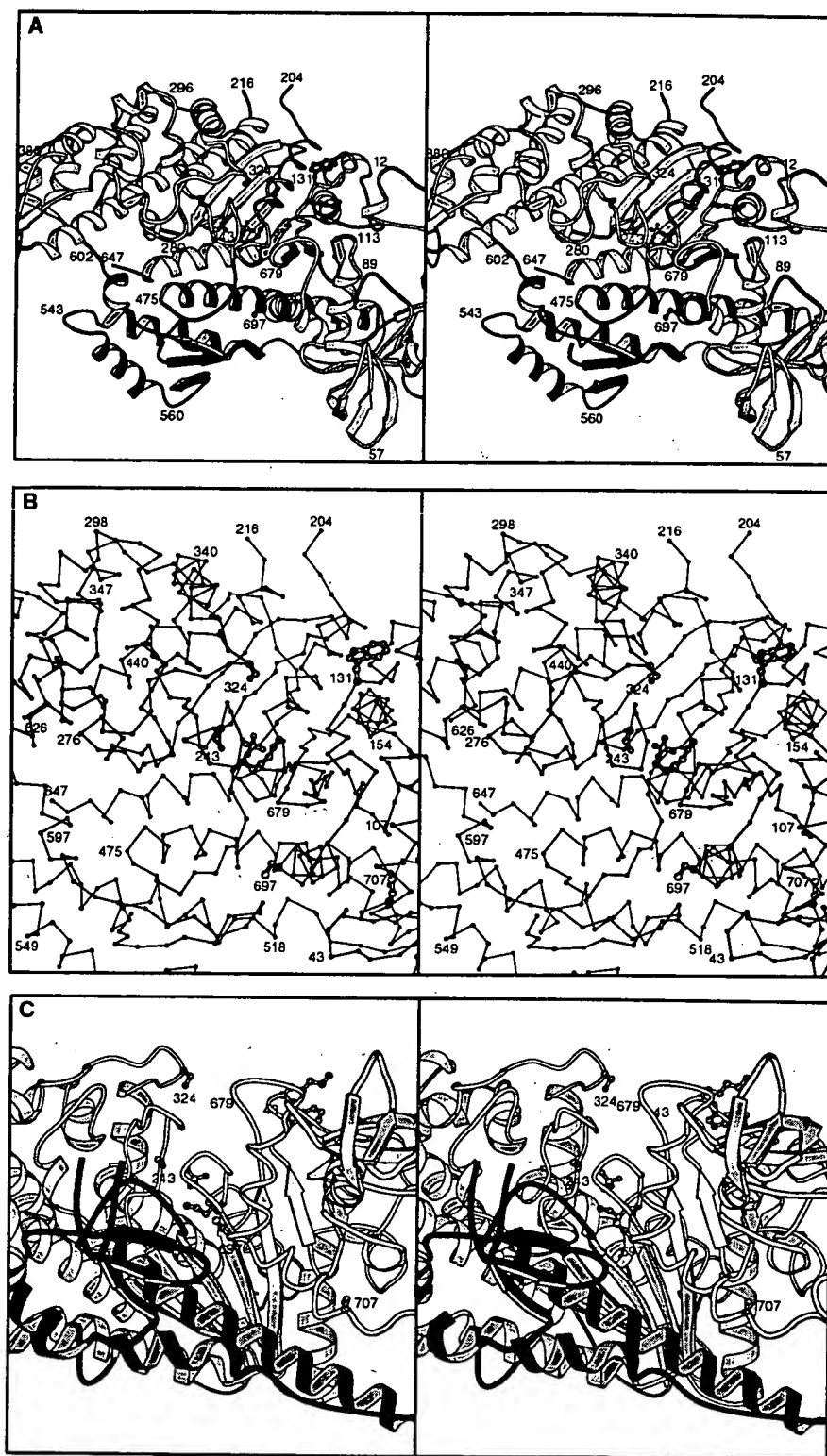


Fig. 6. (A) and (C) show ribbon representations of the regulatory and essential light chains together with the segment of the heavy chain with which they interact. The light chains are oriented such that the NH₂-terminal domains have the same orientation as calmodulin shown in (B). The coordinates for calmodulin were taken from the Brookhaven Protein Data Bank (file 3CLN) from the structure determined by Cook and co-workers (63).

Fig. 7. A stereo ribbon and α carbon plots of the catalytic portion of the myosin head centered on the active site. In (A) and (B) the actin binding face, as defined by the position of the 50- to 20-kD junction, is located on the far side of the molecule. In (C) the molecule has been rotated 90° about the horizontal axis to reveal more clearly the relation between the active site pocket and the reactive cysteine residues. (A) A larger segment of the myosin head that reveals the overall disposition of the secondary structural elements around the nucleotide binding site. The upper domain of the 50-kD segment is shaded in gray whereas the lower domain is shaded in black to emphasize the narrow cleft that divides them. (B) A more detailed view of the residues that form the interface between the upper and lower domains of the 50-kD segment. Marker residues are identified that allow all other residues to be located. In addition, a few of the side chains for the residues that have been implicated to be important in the catalytic mechanism, from amino acid sequence analyses and from chemical studies, have been included. Residues Trp¹³¹ and Ser³²⁴ that have been identified from photolabeling studies lie on opposite sides of the nucleotide binding pocket. (C) The helix connecting the reactive cysteines, Cys⁷⁰⁷ and Cys⁶⁹⁷, lies at the base of a cleft at the junction between the lower domain of the 50-kD segment and the NH₂-terminal 25-kD segment.



tains both the nucleotide binding site and actin binding region. These are located on opposite sides of the protein. This part of the molecule contains a complex arrangement of secondary structural elements centered mainly around a large, mostly parallel, seven-stranded β sheet motif. The topology of this β sheet is such that strands one and six run in the opposite direction to the other five strands. The central strand corresponds to the strand-loop-helix binding motif, which has the sequence GES-GAGKT (39), observed both in adenylate kinase and the Ras protein (40). The topology and organization of the heavy chain are described below in terms of the three major tryptic fragments. However, these fragments arise from proteolytic cleavage at flexible loops and do not represent discrete structural domains.

The first observed residue at the NH₂-terminus of the heavy chain is Asp⁴ and is located close to the essential light chain at the approximate center of the entire myosin molecule (Figs. 4 and 5). From here the heavy chain crosses the width of the molecule and forms a small six-stranded antiparallel β sheet motif (Lys³⁵ to Met⁸⁰), which is fairly independent of the rest of the head and protrudes from the molecule as a whole. The function of this domain is unknown although it does not appear essential for motility in that it is missing in several single-headed myosin I-type molecules (41). The topology of this sheet is similar to that of the Src-homology 3 domain observed in spectrin (42). After this motif, the heavy chain forms three strands of the large β sheet motif that are connected by a series of α helices. The first two strands

extend from Tyr¹¹⁶ to Tyr¹¹⁸ and from Cys¹²³ to Val¹²⁶ and are connected by a β turn. Thereafter there are three short helices prior to the fourth β strand in the sheet that extends from Gln¹⁷³ to Gly¹⁷⁹. The third strand belongs to the COOH-terminal 20-kD fragment of the heavy chain fragment. The fourth or central strand precedes the phosphate binding loop and is followed by a helix,

Lys¹⁸⁵ to Ile¹⁹⁹, which forms the base of the nucleotide binding pocket. The topology of this loop is essentially identical to that in the Ras protein and adenylate kinase (40). A sulfate ion is embedded in the phosphate binding loop and is located close to the position of the β phosphate observed in the complex between Ap₅A [P¹,P⁵,bis-(adenosine-5'-) pentaphosphate] and adeny-

ate kinase (Figs. 4 and 5). It is perhaps not surprising to find a sulfate ion in the nucleotide binding site because ammonium sulfate is a competitive inhibitor of the ATPase activity (43). A break in the electron density is observed between Glu²⁰⁴ and Gly²¹⁶ at the far end of the active site pocket. The missing segment, which contains six charged residues, occurs at the 25- to 50-kD junction and is most likely a constitutively flexible loop.

The 50-kD fragment has a complex topology that can be described as two major domains separated by a long narrow cleft as is evident in the space-filling drawing (Fig. 3). This cleft divides the distal one-third of the myosin head into two regions, which are referred to as the upper and lower domains of the 50-kD segment (Fig. 3).

Electron density for the polypeptide chain resumes at Gly²¹⁶ as the start of an α helix (Leu²¹⁸ to Gly²³³). This helix forms part of the nucleotide binding pocket. Thereafter, the chain loops around close to the phosphate binding site and connects up to β strands six and seven of the large β sheet motif that extend from Gly²⁴⁷ to His²⁵⁴ and Leu²⁶⁰ to Tyr²⁶⁸, respectively. Strand seven terminates in a domain composed of random coil and several short helices and extending from Glu²⁷¹ to Asp³²⁷. This region is located close to the nucleotide binding site and contains Ser³²⁴ which had been previously identified by photolabeling to be an active site residue (44) (Fig. 7). An α helix extending from Asp³²⁷ to Ile³⁴⁰ forms the top of the nucleotide binding pocket. After this domain, the polypeptide chain forms the end of the myosin head through a series of long α helices. The longest of these is 45 Å in length and extends from Val⁴¹⁹ to Leu⁴⁴⁹. Strand five of the large mixed β sheet follows this helix and extends from Tyr⁴⁵⁷ to Ala⁴⁶⁵. This strand terminates in a random coil that drops from the "upper" to "lower" domains of the 50-kD fragment. The midpoint between the upper and lower domains is located close to Gly⁴⁶⁶ and occurs in a region of the sequence (Tyr⁴⁵⁷ to Gly⁵¹⁶) that is highly conserved in all myosins (45). Furthermore, the cleft itself contains many individual highly conserved residues that extend into the space between the two domains.

The lower domain is built from several long α helices (Phe⁴⁷⁵ to Lys 505 and Met⁵¹⁷ to Glu⁵³⁹), the last of which contains a hydrophobic bulge at Pro⁵²⁹. After another helix (Asp⁵⁴⁷ to His⁵⁵⁸) there is a three-stranded antiparallel β sheet, which includes residues Asn⁵⁶⁴ to Lys⁵⁶⁷, Phe⁵⁷⁹ to Val⁵⁸², and Thr⁵⁸⁷ to Tyr⁵⁹⁰. The electron density for Lys⁵⁷², Gly⁵⁷³, and Lys⁵⁷⁴ is very weak, and therefore these residues have been excluded from the model. The segment between Pro⁵²⁹ and Lys⁵⁵³ is one component of the actin binding surface as defined by Ray-

ment *et al.* (46). A single segment of random coil (Lys⁶⁰⁰ to Leu⁶⁰³) passes from the lower domain and across the cleft to form a helix-loop-helix motif on the outer face of the upper 50-kD domain and terminating at Tyr⁶²⁶. There is no electron density corresponding to amino acid residues Gly⁶²⁷ to Phe⁶⁴⁶. This particular stretch contains the second major site of trypsin proteolysis and is the junction between the 50- and 20-kD fragments. The primary sequence in this disordered region contains nine glycine and five lysine residues, suggesting that it may be a flexible region in the molecule. This site is resistant to proteolysis in the actomyosin complex and as such may contribute to the actin binding interface of myosin (33). In addition, this region has also been implicated in actin binding from crosslinking and kinetic studies of proteolytically cleaved protein (34, 47).

Electron density for the polypeptide chain resumes at Gln⁶⁴⁷ and proceeds as a long α helix (Ser⁶⁵⁰ to Arg⁶⁶⁵) across the flat face of the molecule toward the light chain binding region and lies between the upper and lower domains of the 50-kD fragment. This helix is part of a highly conserved segment that runs from Leu⁶⁵⁸ to Asn⁶⁷⁸. At the end of the helix, the polypeptide chain turns into the center of the molecule and forms the third strand of the mixed β sheet (His⁶⁶⁸ to Ile⁶⁷⁵). Thus the major tertiary motif of the head contains contributions from all three of the tryptic fragments. After leaving the β sheet, the polypeptide chain proceeds through the large surface loop defined by Thr⁶⁶⁷ to Glu⁶⁸⁷, which caps one end of the nucleotide binding site pocket. Subsequently, the polypeptide chain forms two α helices lying under the nucleotide binding site and delineated by His⁶⁸⁸ to Asn⁶⁹⁸ and Val⁷⁰⁰ to Arg⁷⁰⁸. This highly conserved segment in the sequence contains the two sulfhydryl groups, Cys⁷⁰⁷ and Cys⁶⁹⁷, which are more reactive than the other 11 in the molecule and have been given the names SH1 and SH2, respectively, in the order of their chemical reactivity. These two thiols can be crosslinked by oxidation and a wide variety of bifunctional chemical reagents differing in length from 14 to 3 Å but only in the presence of nucleotide (48). Indeed, formation of a covalent link between these two groups serves to trap Mg²⁺-ADP (adenosine diphosphate) in the active site. Although these reactive sulfhydryls have been thought to reside in a flexible loop, the discovery that these two residues are separated by an α helix was surprising (Fig. 7C). This is a well-defined region of the electron density map. The fact that the α carbons of Cys⁶⁹⁷ and Cys⁷⁰⁷ are approximately 18 Å apart suggests that a rearrangement or conformational change in this area must occur upon nucleotide bind-

ing. This point is further emphasized by the observation that SH1 and SH2 both lie in small clefts that face out toward the solvent on opposite sides of the molecule. The functional significance of this region is also indicated by the very high degree of amino acid sequence conservation in this area of the molecule.

The segment that follows the reactive sulfhydryl groups consists of a small three-stranded antiparallel β sheet that includes residues Arg⁷¹⁴ to Tyr⁷¹⁷, Tyr⁷⁵⁸ to Gly⁷⁶¹, and Lys⁷⁶⁴ to Phe⁷⁶⁷, and is associated with two short helices. This domain is separated from the adjacent NH₂-terminal domain of the 25-kD fragment of the heavy chain by a distinct cleft and shows a greater association with the COOH-terminal domain of the essential light chain. Thereafter the heavy chain continues as a long α helix that shows distinct curvature beginning at Leu⁷⁷¹ and ending at Val⁸²⁶. There is a decided bend in the course of the polypeptide chain resulting from the Trp⁸²⁹, Pro⁸³⁰, Trp⁸³¹ sequence (Figs. 4 and 5). The heavy chain terminates at residue Lys⁸⁴³ after a small α helix that lies nearly at right angles to the preceding long helix.

Effect of the reductive methylation on the protein structure and function. One question that must be addressed is the effect, if any, of reductive methylation on the conformation of the protein. An examination of the kinetic properties of modified myosin S1 reveals that the protein is enzymatically active (49). There are changes in the kinetic parameters that are similar to those observed when only the reactive sulfhydryl groups are alkylated (50). The results do not suggest any major changes in the overall conformation of the molecule since these would be expected to abolish its enzymatic activity. Myosin from most sources already contains several post-translationally modified amino acid residues. For example, in chicken skeletal myosin S1, Lys³⁵ is monomethylated, Lys¹³⁰ and Lys⁵⁵¹ are trimethylated, and His⁷⁵⁷ contains a 3-N-methylated side chain (27). Although the role of these modified residues is unknown, it has been suggested that methylation of Lys¹³⁰ provides a permanent positive charge that may become buried when nucleotide is bound (51). In our structure, Lys¹³⁰ is exposed to the solvent at the edge of the nucleotide binding pocket. However, this region of the protein probably rearranges when nucleotide binds because the adjacent Trp¹³¹ is photolabeled by two purine ATP analogues (52).

The structure of methylated lysozyme is essentially identical to that of the native protein (15). From this it is not expected that the folding motifs in myosin S1 will be significantly affected by this treatment. There is, however, the possibility that the relation between the various domains could be altered. Our data reveal that almost all of

the lysine residues are located at the surface and hence would not be expected to influence the structure in any major way. The A2 isozyme of chicken myosin S1 contains 102 lysine residues (27), of which 85 have been built into the model. Of those, 67 are located at the surface of the protein and only 18 participate in salt bridges and can be considered buried. Five of these lysines participate in crystalline contacts. The remaining 17 lysine residues in the A2 isozyme are located in disordered loops; in our structure all except 4 lysine residues are reproducibly modified under conditions where 100 percent dimethylation is expected (Table 1) (15). Thus, the unmodified lysine residues are most likely located in salt bridges where they would be expected to have a higher pK_a . Mass for the additional methyl groups on the lysine residues is evident in the electron density map for most of the well-ordered side chains. However at this resolution it is difficult to categorically decide if a residue has been modified based on the density alone. Even so, it appears that Lys¹⁸⁵, which resides in the phosphate binding loop, is not modified.

Active site and possible mechanism for muscle contraction. The catalytic site of the myosin head was identified by analogy to the phosphate binding loop in both the Ras protein and adenylate kinase and by the position of the amino acid residues previously identified by chemical studies with ATP analogues (52). The nucleotide binding pocket is located on the opposite side of the head from the proposed actin binding site and is in an open conformation (Figs. 4, 5, and 7). The view in Fig. 7 shows the position of the sulfate ion in the phosphate binding loop and a few of the amino acid residues that have been chemically labeled, including Trp¹³¹, Ser¹⁸¹, Ser²⁴³, and Ser³²⁴ (44, 52, 53). The width of the nucleotide binding pocket at its surface is approximately 15 Å as measured between α carbons. Since the binding constant of myosin for Mg^{2+} -ATP is about 3×10^{11} (54) and residues on both sides of the cleft have been photochemically labeled, it is likely that the pocket closes when nucleotides bind in the active site. The pocket is approximately 13 Å wide and 13 Å deep with an angle between the faces of the pocket of $\sim 40^\circ$. The base of the cleft is located 90 Å from the COOH-terminus of the myosin head. If the binding face to actin remains essentially stationary, closure of the nucleotide binding cleft could produce a movement at the COOH-terminus of the myosin head of approximately 60 Å. How this rearrangement is actually accomplished cannot be easily predicted from our structure.

The orientation of the molecule in Fig. 5 is rotated such that the actin binding surface is approximately perpendicular to the page (46). Closure of the nucleotide

binding pocket would rotate the COOH-terminal end of the heavy chain that carries the light chains toward the viewer, which is consistent with that expected for the start of the power stroke. From this perspective it appears that a major function of the light chains is to create a longer molecule and hence amplify the conformational changes associated with the active site.

Muscle contraction consists of the cyclic attachment and detachment of the myosin head to the actin filament with the concomitant hydrolysis of ATP. From the extensive kinetic studies on the interaction of myosin with actin (55), a general picture of the sequence of kinetic events occurring during muscle contraction has emerged.

Transient kinetic measurements originally demonstrated that transduction of the chemical energy released by the hydrolysis of ATP into directed mechanical force occurred during product release rather than during the hydrolysis step itself (56). The cycle of events was summarized as follows: Mg^{2+} -ATP rapidly dissociates the actomyosin complex by binding to the ATPase sites of myosin; free myosin then hydrolyzes ATP and forms a relatively stable myosin-products complex; actin recombines with this complex and dissociates the products, thereby forming the original actin-myosin complex. Presumably, force is generated during the last step. Although this model provided an important conceptual framework for studies of the contractile cycle, it soon became clear that the interactions between myosin, actin, and the substrate and products were more complex (55).

Structural information on the conformational changes that occur during the actomyosin interactions is limited. Addition of ATP causes no significant change in the amount of secondary structure as assessed by circular dichroism (57). The changes observed in tryptophan fluorescence are typical of most enzymes whose active sites are induced to fit around their substrates. However, significant movement within the myosin head must occur during the ATPase activity because of the large change in distance between the two reactive cysteine residues (Cys⁷⁰⁷ and Cys⁶⁹⁷) that is induced when nucleotide binds (48, 58). Recent low-angle x-ray scattering studies also suggest a large-scale movement during ATP hydrolysis (59).

In formulating a model for muscle contraction from the structure of myosin S1 presented here, it must be understood that it neither contains nucleotide nor is bound to actin. Most likely the crystal structure is an intermediate between these two extremes, although probably closer to the actin bound state. Preliminary attempts to dock myosin (46) to actin suggest that a better fit to the image reconstructions of S1-decorated actin would be obtained if the long narrow cleft between the upper and lower 50-kD domains were to

close, thus implying that this is an important structural feature of the molecule. In addition, the preliminary fit implies that the actin binding site contains components from both the upper and lower 50-kD domains and the first α helix from the 20-kD region. From the location of residues Tyr⁶²⁶ and Gln⁶⁴⁷, the positively charged disordered segment at the 50- to 20-kD junction could readily interact with the negatively charged amino acids at the NH₂-terminus of actin.

All the current kinetic models for the mechanism of muscle contraction require a change in the binding affinity of myosin for actin when ATP binds to the active site. Although it is difficult to predict how this effect can be communicated to the actin binding site, the structure suggests that this might be generated by changes in the relation between the upper and lower domains of the 50-kD segment prompted by binding of the γ phosphate. Examination of Fig. 7 reveals that the potential binding site for the γ phosphate would be located close to the confluence of the upper and lower domains of the 50-kD region below the current location of the sulfate ion. These observations together with the information from docking myosin onto actin provide the information necessary to formulate a basic structural model for muscle contraction (46).

The three-dimensional model of the myosin S1 presented in this article provides a molecular framework that can be used to address the issues of conformational changes during the contractile cycle and suggests how this molecule functions as a molecular motor. By a combination of molecular biology, in vitro motility assays, and chemical and kinetic studies, it should be possible to test these hypotheses concerning the molecular basis of motility.

Finally, it is appropriate to consider why reductive methylation allows this molecule to crystallize. Examination of the structure reveals that it contains elements of flexibility that might lead to multiple conformations in solution, which in turn might prevent the formation of a crystalline lattice. It is conceivable that reductive methylation serves to stabilize one of these conformations in solution. Alternatively, reductive methylation may serve only to reduce the solubility of the protein.

REFERENCES AND NOTES

1. R. D. Vale and L. S. B. Goldstein, *Cell* 60, 883 (1990).
2. A. F. Huxley and R. Niedergerke, *Nature* 173, 971 (1954); H. Huxley and J. Hanson, *ibid.* p. 973.
3. S. Lowey, H. S. Slayter, A. G. Weeds, H. Baker, *J. Mol. Biol.* 42, 1 (1969).
4. A. G. Weeds and S. Lowey, *ibid.* 61, 701 (1971).
5. Y. Y. Toyoshima *et al.*, *Nature* 328, 536 (1987).
6. S. S. Margossian, W. F. Stafford III, S. Lowey, *Biochemistry* 20, 2151 (1981).
7. P. D. Wagner and E. Giniger, *Nature* 292, 560 (1981).

8. S. Citi and J. Kendrick-Jones, *BioEssays* 7, 155 (1987).
9. J. S. Sellers, *Curr. Opin. Cell Biol.* 1, 98 (1991).
10. J. H. Collins, *J. Muscle Res. Cell Motil.* 12, 3 (1991).
11. S. Lowey, in *Myology*, A. G. Engel and B. Q. Banker, Eds. (McGraw-Hill, New York, 1986), vol. 19, pp. 563-586; P. Vibert and C. Cohen, *J. Muscle Res. Cell Motil.* 9, 296 (1988).
12. A. Elliott and G. Offer, *J. Mol. Biol.* 123, 505 (1978).
13. D. A. Winkelmann, H. Mekeel, I. Rayment, *ibid.* 181, 487 (1985); D. A. Winkelmann, T. S. Baker, I. Rayment, *J. Cell Biol.* 114, 701 (1991).
14. R. H. Rice and G. E. Means, *J. Biol. Chem.* 246, 831 (1971).
15. W. R. Rypniewski, H. M. Holden, I. Rayment, *Biochemistry*, in press.
16. L. Silberstein and S. Lowey, *J. Mol. Biol.* 148, 153 (1981).
17. Y. Nabeshima, Y. Fujii-Kuriyama, M. Muramatsu, K. Ogata, *Nature* 308, 333 (1984).
18. R. Smith, W. R. Rypniewski, I. Rayment, in preparation.
19. I. Rayment and D. A. Winkelmann, *Proc. Natl. Acad. Sci. U.S.A.* 81, 4378 (1984).
20. The data to 3.5 Å resolution were recorded at 4°C on a Siemens X1000D area detector; 13 crystals were used to collect the native data. Most of these crystals were translated to expose a new region after 10 to 11 hours in the x-ray beam such that the data were collected from 35 segments. A total of 94,265 reflections were measured, which reduced to 21,370 unique reflections (Theoretical 24,556) with an R_{merge} of 5.3 where

$$R_{\text{merge}} = \frac{\sum \sum (|I_{hi}| - |I_{li}|) / \sum I_{hi} \times 100}{\sum I_{hi}}$$

$$I_{hi}$$
 and I_{li} are the intensities of the individual and mean structure factors, respectively. These data were recorded with the goal of obtaining a complete, accurate low-resolution x-ray data set that could be used to determine the positions of the heavy atoms in the derivatives. The frame data were processed by the program XDS (64) and scaled with the Fox and Holmes algorithm as implemented by P. Evans in the programs Rotavata and Agrovata (65). The x-ray data between 3.5 and 2.8 Å were recorded on film at the synchrotron sources located at Cornell (CHESS) and Stanford (SSRL). Data were processed with the software developed by M. Rossmann, modified to operate on a VAX (66). The data were merged and scaled with the same software used for the area detector data, except for the inclusion of post-refinement to utilize the partial data (66). A total of 178,986 measurements were recorded on 97 films to yield 36,781 independent reflections in the SSRL native data set with an R_{merge} of 9.2. The final native data set consisted of structure factors from 100 to 4.5 Å recorded on the area detector and data from 4.5 to 2.8 Å recorded at SSRL.
21. H. M. Holden and I. Rayment, *Arch. Biochem. Biophys.* 291, 187 (1991).
22. M. G. Rossmann, *Acta Crystallogr.* 13, 221 (1960); T. C. Terwilliger and D. Eisenberg, *Acta Crystallogr. Sect. A* 39, 813 (1983).
23. W. A. Hendrickson and E. E. Lattman, *ibid.* *Acta Crystallogr. Sect. B* 26, 136 (1970).
24. B. C. Wang, *Methods Enzymol.* 115, 90 (1985). The algorithm was written by W. Kabsch (Heidelberg, Germany).
25. M. A. Rould, J. J. Perona, D. Söhl, T. A. Steitz, *Science* 246, 1135 (1989).
26. T. A. Jones, *Methods Enzymol.* 115, 157 (1985).
27. T. Maita *et al.*, *J. Biochem.* 110, 75 (1991); G. Matsuda, *Adv. Biophys.* 16, 185 (1983).
28. The size of the side chains observed in the electron density map was matched to the sequence with the program FITSEQ, available from I. Rayment on request.
29. R. J. Read, *Acta Crystallogr. Sect. A* 42, 140 (1986).
30. D. E. Tronrud, L. F. Ten Eyck, B. W. Matthew, *ibid.* 43, 489 (1987).
31. A. T. Brunger, *X-PLOR Manual Version 2.1* (Yale University, New Haven, CT, 1990).
32. L. Szilagyi, M. Balint, F. A. Sreter, J. Gergely, *J. Biochem. Biophys. Res. Commun.* 87, 936 (1979).
33. D. Mornet, R. Bertrand, P. Pantel, E. Audemard, R. Kassab, *Biochemistry* 20, 2110 (1981).
34. K. Sutoh, *ibid.* 21, 4800 (1982).
35. Y. S. Babu *et al.*, *Nature* 315, 37 (1985).
36. O. Herzberg and M. N. G. James, *ibid.* 313, 653 (1985).
37. G. Matsuda, Y. Suzuyama, T. Maita, T. Umegane, *FEBS Lett.* 84, 53 (1977).
38. M. Ikura *et al.*, *Science* 256, 632 (1992); W. E. Meador, A. R. Means, F. A. Quirocho, *ibid.* 257, 1251 (1992).
39. Abbreviations for the amino acids residues are A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
40. C. W. Muller and G. E. Schutz, *J. Mol. Biol.* 224, 159 (1992); E. F. Pai *et al.*, *EMBO J.* 9, 2351 (1990).
41. T. D. Pollard, S. K. Doberstein, H. G. Zot, *Annu. Rev. Physiol.* 53, 653 (1991).
42. A. Musacchio *et al.*, *Nature* 359, 851 (1992).
43. C. Tesi, T. Barman, F. Travers, *FEBS Lett.* 236, 256 (1988).
44. R. Mahmood, M. Elzinga, R. G. Yount, *Biochemistry* 28, 3989 (1989).
45. H. M. Warrick and J. A. Spudich, *Annu. Rev. Cell Biol.* 3, 379 (1987).
46. I. Rayment *et al.*, *Science* 261, 58 (1993).
47. K. Yamamoto, *J. Mol. Biol.* 217, 229 (1991).
48. M. Burke and E. Reisler, *Biochemistry* 16, 5559 (1977); J. A. Wells and R. G. Yount, *Methods Enzymol.* 85, 93 (1982).
49. H. White and I. Rayment, *Biochemistry*, in press.
50. J. A. Sleep, K. M. Trybus, K. A. Johnson, E. W. Taylor, *J. Muscle Res. Cell Motil.* 2, 373 (1981).
51. Y. Okamoto and R. G. Yount, *Proc. Natl. Acad. Sci. U.S.A.* 82, 1575 (1985).
52. R. G. Yount, C. R. Cremon, J. C. Grammer, B. A. Kerwin, *Philos. Trans. R. Soc. London Ser. B* 336, 55 (1992).
53. C. R. Cremon, J. C. Grammer, R. G. Yount, *J. Biol. Chem.* 264, 6608 (1989); J. C. Grammer and R. G. Yount, *Biophys. J.* 59, 226a (1991).
54. R. S. Goody, W. Hofmann, H. G. Mannherz, *Eur. J. Biochem.* 78, 317 (1977).
55. E. W. Taylor, *CRC Crit. Rev. Biochem. Biophys.* 6, 103 (1979); R. S. Adelstein and E. Eisenberg, *Annu. Rev. Biochem.* 49, 921 (1980); M. G. Hibberd and D. R. Trentham, *Annu. Rev. Biophys. Biochem.* 15, 119 (1986); M. A. Geeves, *Biochem. J.* 274, 1 (1991).
56. R. W. Lynn and E. W. Taylor, *Biochemistry* 10, 4617 (1971).
57. W. B. Gratzer and S. Lowey, *J. Biol. Chem.* 244, 22 (1969).
58. E. E. Huston, J. C. Grammer, R. G. Yount, *Biochemistry* 27, 8945 (1988).
59. K. Wakabayashi *et al.*, *Science* 258, 443 (1992).
60. Amino acid analyses were performed by L. Mende-Mueller at the Protein and Nucleic Acid Facility, Medical College of Wisconsin, Milwaukee, WI 53226.
61. T. E. Ferrin *et al.*, *J. Mol. Graphics* 6, 13 (1988).
62. P. J. Kraulis, *J. Appl. Cryst.* 24, 946 (1991).
63. Y. S. Babu, C. E. Bugg, W. J. Cook, *J. Mol. Biol.* 204, 191 (1988).
64. W. Kabsch, *J. Appl. Cryst.* 21, 67 (1988); *ibid.*, p. 916.
65. G. C. Fox and K. C. Holmes, *Acta Crystallogr.* 20, 886 (1966).
66. M. G. Rossmann, *Methods Enzymol.* 114, 237 (1985).
67. We thank the co-directors (P. A. Frey, W. W. Cleland, H. Lardy, and G. H. Reed) at the Institute for Enzyme Research for support and discussion, K. Johnson and J. Dewane for technical assistance, and J. Sakon and J. Wedekind for help in the synchrotron data collection and processing. This project was initiated at Brandeis University and continued at the University of Arizona; we thank D. L. D. Caspar, J. H. Law, and M. A. Wells at those institutions for their support and encouragement. Supported by NIH grants (I.R., H.M.H., and D.A.W.). I.R. thanks R. Yount for support and encouragement during the long process of solving this structure.

6 April 1993; accepted 4 June 1993

Structure of the Actin-Myosin Complex and Its Implications for Muscle Contraction

Ivan Rayment,* Hazel M. Holden, Michael Whittaker, Christopher B. Yohn, Michael Lorenz, Kenneth C. Holmes, Ronald A. Milligan

Muscle contraction consists of a cyclical interaction between myosin and actin driven by the concomitant hydrolysis of adenosine triphosphate (ATP). A model for the rigor complex of F actin and the myosin head was obtained by combining the molecular structures of the individual proteins with the low-resolution electron density maps of the complex derived by cryo-electron microscopy and image analysis. The spatial relation between the ATP binding pocket on myosin and the major contact area on actin suggests a working hypothesis for the crossbridge cycle that is consistent with previous independent structural and biochemical studies.

Muscle contraction occurs when two sets of interdigitating filaments, the thin actin filaments and the thick myosin filaments, slide past one another. A widely accepted theory to explain this process is the cross-bridge hypothesis of muscle contraction whereby sliding is brought about by cross-

bridges that extend from the myosin filament and interact cyclically in a rowing motion with the actin filament as adenosine triphosphate (ATP) is hydrolyzed (1, 2).

The myosin head is an actin-activated adenosine triphosphatase (ATPase). Both solution kinetic studies and fiber experi-

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER^{*†‡}, CYRUS CHOTHIA^{*}, AND TIM J. P. HUBBARD[§]

^{*}MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and [†]Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA $ktup = 1$, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith-Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ($ktup = 2$) or greater effectiveness ($ktup = 1$). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPO, errors per query.

[†]Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

[‡]To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

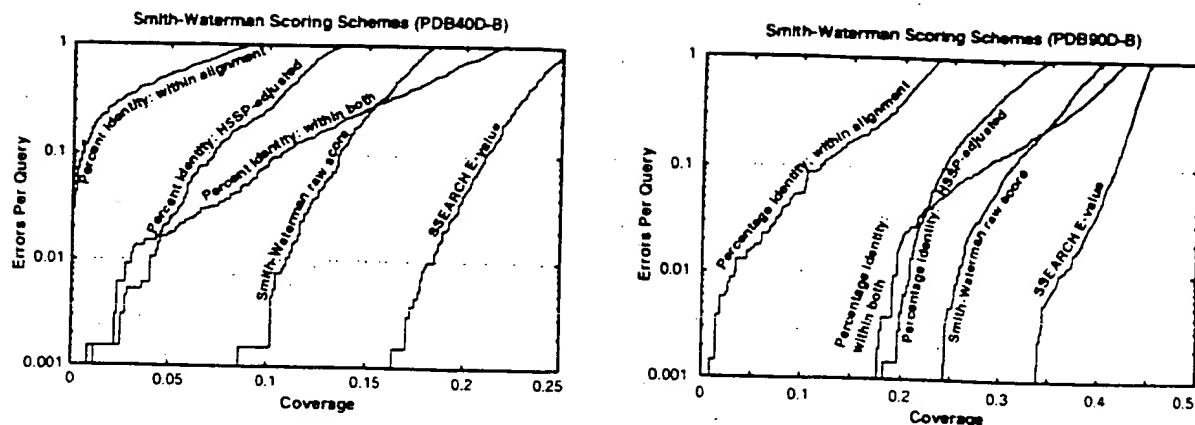
From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ~0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0i76 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties -12/-1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have



perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

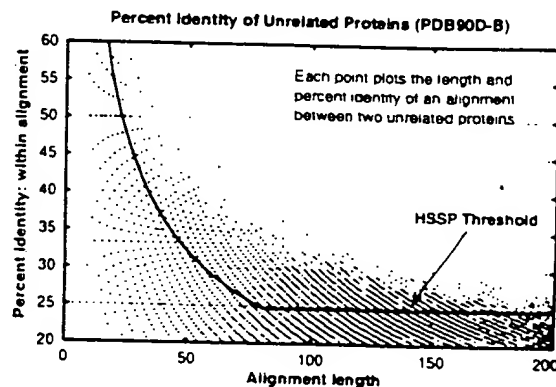
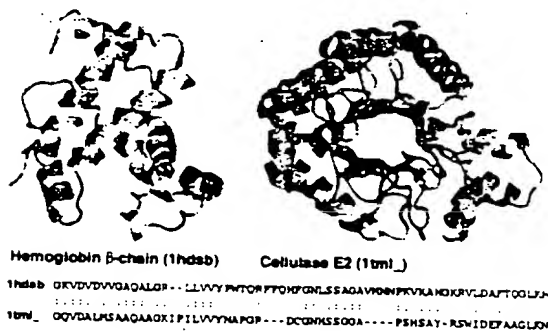


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSP threshold (though it is intended to be applied with a different matrix and parameters).

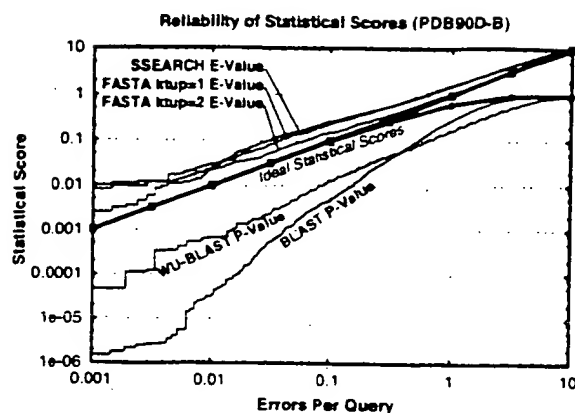


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

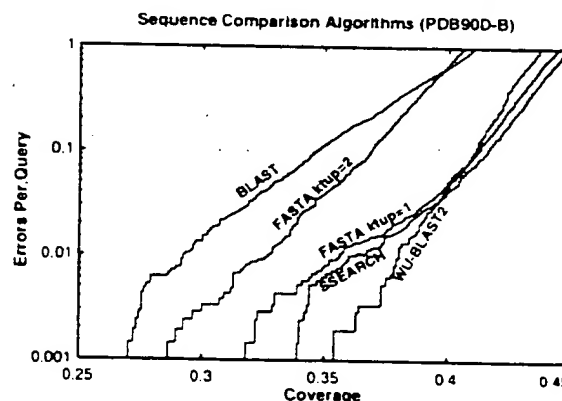
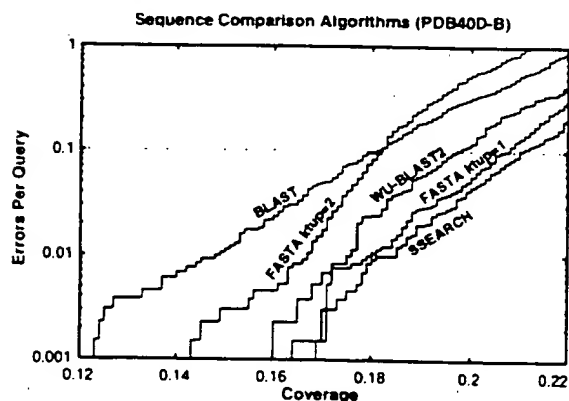


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPO for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPO.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPO. BLAST, which identifies 15%, was the worst performer, whereas FASTA $k_{\text{up}} = 1$ is nearly as effective as SSEARCH. FASTA $k_{\text{up}} = 2$ and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA $k_{\text{up}} = 1$. WU-BLAST2 is slightly faster than FASTA $k_{\text{up}} = 2$, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA $k_{\text{up}} = 1$, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

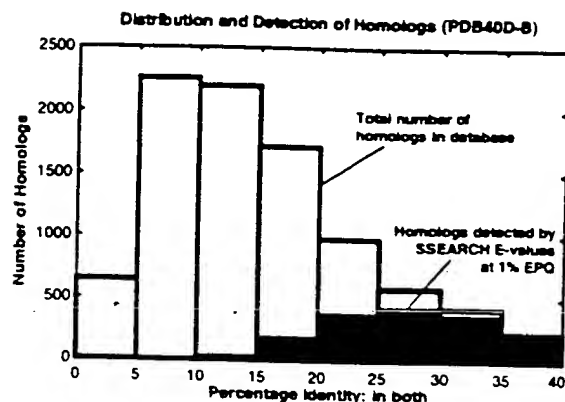


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPO. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPO Cutoff	Coverage at 1% EPO
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSP-scaled	25.5	35% (HSP = 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{up}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{up}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403-410.
- Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460-480.
- Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444-2448.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* 247, 536-540.
- Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* 266, 635-643.
- Pearson, W. R. (1991) *Genomics* 11, 635-650.
- Pearson, W. R. (1995) *Protein Sci.* 4, 1145-1160.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195-197.
- George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* 266, 41-59.
- Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* 249, 816-831.
- Henikoff, S. & Henikoff, J. G. (1993) *Proteins* 17, 49-61.
- Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* 24, 21-25.
- Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* 24, 189-196.
- Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* 89, 10915-10919.
- Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345-352.
- Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
- Sander, C. & Schneider, R. (1991) *Proteins* 9, 56-68.
- Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* 233, 716-738.
- Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* 1, 89-94.
- Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* 1, 77-78.
- Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* 14, 971-993.
- Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* 87, 2264-2268.
- Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* 90, 5873-5877.
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* 6, 119-129.
- Pearson, W. R. (1996) *Methods Enzymol.* 266, 227-258.
- Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* 12, 215-226.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* 266, 554-571.
- Waterman, M. S. & Vingron, M. (1994) *Stat. Science* 9, 367-381.
- Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* 13, 669-678.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107-132.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* 7, 369-376.
- Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997) *Structure (London)* 5, 1093-1108.
- Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* 39, 561-577.
- Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* 20, 25-33.
- Fitch, W. M. (1966) *J. Mol. Biol.* 16, 9-16.
- Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* 4, 1123-1127.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389-3402.
- Girling, R., Schmidt, W., Jr., Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* 131, 417-433.
- Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* 32, 9906-9916.
- Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374-376.

hmmpfam - search a single seq against HMM database
 HMMER 2.1.1 (Dec 1998)
 Copyright (C) 1992-1998 Washington University School of Medicine
 HMMER is freely distributed under the GNU General Public License (GPL).

HMM file: /data/isb2k/blastdb/Pfam72/Pfam72
 Sequence file: /u/legal/jennyb/pf621.seq

Query: 1929760CD1

Scores for sequence family classification (score includes all domains):

Model	Description	Score	E-value	N
myosin_head	Myosin head (motor domain)	-188.6	1.2e-19	1

Parsed for domains:

Model	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
myosin_head	1/1	5	388	1	734	-188.6	1.2e-19

Alignments of top-scoring domains:

myosin_head: domain 1 of 1, from 5 to 388: score -188.6, E = 1.2e-19

```

      *->vEDmveLtyLnEpsvlhNLKkRYksdliITYsGlvLvsvNPYkrLpq
1929760CD1    5  -----p- 5
                  iYteeiiakYrGKrryElPPHiFAiAdeAYRsMlsdkeNQsilISGESGA
1929760CD1    -  -----
                  GKTEntKkvmqYlAaVsggnsngngeevpskvgrvEdqILqsNPiLEAFG
                  Vs++ s
1929760CD1    6  -----QVSCSLs----- 12
                  NAKTtRNNNSSRFGKyieIqFdktGkivGakIenYLLEKSRVvyQteger
1929760CD1    -  -----
                  NFHIFYQLLaGasqqnlkkeLkLtndpedYhYLnqgggevkpcytvdGiDD
                  +++      L  +++      qg+
1929760CD1    13 -----LMPR-----LP-SIRHW-----QGP----- 26
                  segnveeFketrkAmdilGftdeeqrsIFrivAaILhlGNikFkqrrkee
                  s
1929760CD1    27 S----- 27
                  aaipddnnadtkalekaaeLlGvdatelekALLsrriktGtegrkStvtk
                  +++      ++      G  +
1929760CD1    28 ---HPG-----FL-----GPLF----- 36
                  pqnveQAsyARDALAKalySRlFdWIVnrINktLdfkakegqdasfIGVL
                  p      +L+  +++      a f G L
1929760CD1    37 PI-----CSLQWPHGFS--AIFPGLL 55
                  DIyGFEIFekNSFEQLCINyVNEKLQQfFNhnmFklEQEEYkrEGIEwtf
                  D yGFE F +NS EQLCINy+NEKLQQ+F h + + QEEY EG ew+f
1929760CD1    56 DVYGFESFPDnsLEQLCINyANEKLQQHFVAHYLRAQQEEYAVEGLEWSF 105
                  IdFgdNLQpcIDLIEkKsPpGILsLLDEeClfPkaqSGtDqtFldKLyst
                  I++ dN Qpc DLIE+ P+ I sL+ EeC++ + + + + +
1929760CD1    106 INYQDN-QPCLDLIEGS-PISICSLINEECRLNRPS--SARQLQTRIETA 151
                  fskhpahfekfsPrfrqkksghFiikHYAGdVeYnvegFleKNKDpLfd
                  + p +      +++++ s Fi++HYAG V+Y + g +eKNKDp+++
1929760CD1    152 LAGSPCLGHN---KLSREPS---FIVVHYAGPVRyHTAGLVEKNKDPIPP 195
  
```

17. I declare further that all statements made herein of my own knowledge are true and that all statements made herein on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, and that willful false statements may jeopardize the validity of this application and any patent issuing thereon.

Tod Bedilion

Signed at Redwood City, California

this ____ day of July, 2003

```

dli1lksSsnpllaeLFpdeetlagpfeadpsslskkrksgskNkstgk
+l 1l++S++pll++LFp p++++++ +
1929760CD1 196 ELTRLQSQDPLLMGLFP-----TNPKEKTQEE-----P 225

ktkksnfi.TvGaqlKeslneLMktLsstnLPHFvRCIkPNekKkagvfD
++++ + Tv ++fK sl L+ L st+ PH++RCIkPN+ +a +f
1929760CD1 226 PGQSRAPVlTVVSKFKASLEQLLQVLHSTT-PHYIRCIKPNsQGQAQTFL 274

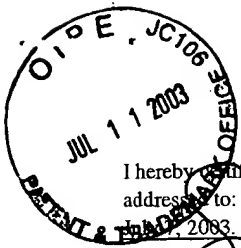
aslVlhQLrclGVLEgiRirRaGFPnRitfdeFlqRYriLapktwP....
++ Vl+QL ++G+ E+i I+ aGFP R+ + F++RY++L + +++++
1929760CD1 275 QEEVLSQLEACGLVETIHISAAGFPiRVSHRNfVERYKLLRRLHPctssg 324

.....kwsgdakkeknEIvaceklLqsLn.....
++++ + ++ ++w +++ + e l+q+ ++ + ++ ++
1929760CD1 325 pdspypakglpEWCPHSEEA-----TLEPLIQDILhtlpvltqaaaitg 368

.....lDkgeeyrfGkTKIFFR<-*
++ + + + + G TK+F
1929760CD1 369 dsaeaMPA-PMH-CGRTKVFMT 388

```

//



Docket No.: PF-0621 USN

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop: Non-Fee Amendment, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on JUL 11 2003.

By: _____

Printed: Lyza Fimiliar

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

RECEIVED
JUL 15 2003
TECH CENTER 1600/2900

In re Application of: Tang et al.

Title: MYOSIN HEAVY CHAIN HOMOLOG

Serial No.: 09/830,914

Filing Date: May 01, 2001

Examiner: Fronda, C.

Group Art Unit: 1652

Mail Stop: Non-Fee Amendment
Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

**DECLARATION OF DR. TOD BEDILION
UNDER 37 C.F.R. § 1.132**

I, TOD BEDILION, a citizen of the United States, residing at 132 Winding Way, San Carlos, California, declare that:

1. I was employed by Incyte Genomics, Inc. (hereinafter "Incyte") as a Director of Corporate Development until May 11, 2001. I am currently under contract to be a Consultant to Incyte Genomics, Inc.

2. In 1996, I received a Ph.D. degree in Cell, Molecular and Development Biology from UCLA. I had previously received, in 1988, a B.S. degree in biology from UCLA.

Upon my graduation from UCLA, I became, in April 1996, the first employee of Synteni, Inc. (hereinafter "Synteni"). I was a Research Director at Synteni from April 1996 until Synteni was acquired by Incyte in early 1998.

I understand that Synteni was founded in 1994 by T. Dari Shalon while he was a graduate student at Stanford University. I further understand that Synteni was founded for the purpose

of commercially exploiting certain "cDNA microarray" technology that was being worked on at Stanford in the early to mid-1990s. That technology, which I will sometimes refer to herein as the "Stanford-developed cDNA microarray technology", was the subject of Dr. Shalon's doctoral thesis at Stanford. I understand and believe that Dr. P.O. Brown was Dr. Shalon's thesis advisor at Stanford.

During the period beginning before I was employed by Synteni and ending upon its acquisition by Incyte in early 1998, I understand Synteni was the exclusive licensee of the Stanford-developed cDNA microarray technology, subject to any right that the United States government may have with respect to that technology. In early 1998, I understand Incyte acquired rights under the Stanford-developed cDNA microarray technology as part of its acquisition of Synteni.

I understand that at the time of the commencement of my employment at Synteni in April 1996, Synteni's rights with respect to the Stanford-developed cDNA technology included rights under a United States patent application that had been filed June 7, 1995 in the names of Drs. Brown and Shalon and that subsequently issued as United States Patent No. 5,807,522 (the Brown '522 patent). In December 1995, the subject matter of the Brown '522 patent was published based on a PCT patent application that had also been filed in June 1995. The Brown '522 patent (and its corresponding PCT application) describes the use of the Stanford-developed cDNA technology in a number of gene expression monitoring applications, as will be discussed more fully below.

Upon Incyte's acquisition of Synteni, I became employed by Incyte. From early 1998 until late 1999, I was an Associate Research Director at Incyte. In late 1999, I was promoted to the position of Director, Corporate Development.

I have been aware of the Stanford-developed cDNA microarray technology since shortly before I commenced my employment at Synteni. While I was employed by Synteni, virtually all (if not all) of my work efforts (as well as the work efforts of others employed by Synteni) were directed to the further development and commercial exploitation of that cDNA microarray technology. By the end of 1997, those efforts had progressed to the point that I understand Incyte agreed to pay at least about \$80 million to acquire Synteni. Since I have been employed by Incyte, I have continued to work

on the further development and commercial exploitation of the cDNA microarray technology that was first developed at Stanford in the early to mid-1990s.

3. I have reviewed the specification of a United States patent application that I understand was filed on May 2, 2001 in the names of Tang et al. and was assigned Serial No. 09/830,914 (hereinafter “the Tang ‘914 application”). Furthermore, I understand that this United States patent application claimed priority to United States Provisional Patent Application Serial No. 60/172,248 filed on November 5, 1998 (hereinafter “the Tang ‘248 application”). The SEQ ID NO:1-encoding polynucleotides were described in the Tang ‘248 application. My remarks herein will therefore be directed to the Tang ‘248 patent application, and November 5, 1998, as the relevant date of filing. In broad overview, the Tang ‘248 specification pertains to certain nucleotide and amino acid sequences and their use in a number of applications, including gene expression monitoring applications that are useful in connection with (a) developing drugs (e.g., the diagnosis of inherited and acquired genetic disorders, expression profiling, toxicology testing, and drug development with respect to cancer, an immunopathology, a neuropathology, and the like), and (b) monitoring the activity of drugs for purposes relating to evaluating their efficacy and toxicity.

4. I understand that (a) the Tang ‘248 application contains claims that are directed to isolated and purified polynucleotides having the sequences disclosed in the Tang ‘914 application as SEQ ID NO:1-encoding polynucleotides, for example SEQ ID NO:2 (hereinafter “the SEQ ID NO:1-encoding polynucleotides”), and (b) the Patent Examiner has rejected those claims on the grounds that the specification of the Tang ‘248 application does not disclose a substantial, specific and credible utility for the claimed SEQ ID NO:1-encoding polynucleotides. I further understand that whether or not a patent specification discloses a substantial, specific and credible utility for its claimed subject matter is properly determined from the perspective of a person skilled in the art to which the specification pertains at the time of the patent application was filed. In addition, I understand that a substantial, specific and credible utility under the patent laws must be a “real-world” utility.

5. I have been asked (a) to consider with a view to reaching a conclusion (or conclusions) as to whether or not I agree with the Patent Examiner's position that the Tang '248 application does not disclose a substantial, specific and credible "real-world" utility for the claimed SEQ ID NO:1-encoding polynucleotides, and (b) to state and explain the bases for any conclusions I reach. I have been informed that, in connection with my considerations, I should determine whether or not a person skilled in the art to which the Tang '248 application pertains on November 5, 1998 would have concluded that the Tang '248 application disclosed, for the benefit of the public, a specific beneficial use of the SEQ ID NO:1-encoding polynucleotides in their then available and disclosed form. I have also been informed that, with respect to the "real-world" utility requirement, the Patent and Trademark Office instructs its Patent Examiners in Section 2107 of the Manual of Patent Examining Procedure, under the heading "I. 'Real-World Value' Requirement":

"Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact 'useful' in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm."

6. I have considered the matters set forth in paragraph 5 of this Declaration and have concluded that, contrary to the position I understand the Patent Examiner has taken, the specification of the Tang '248 patent application disclosed to a person skilled in the art at the time of its filing a number of substantial, specific and credible real-world utilities for the claimed SEQ ID NO:1-encoding polynucleotides. More specifically, persons skilled in the art on November 5, 1998 would have understood the Tang '248 application to disclose the use of the SEQ ID NO:1-encoding polynucleotides in a number of gene expression monitoring applications that were well-known at that time to be useful in connection with the development of drugs and the monitoring of the activity of such drugs. I explain the bases for reaching my conclusion in this regard in paragraphs 7-16 below.

7. In reaching the conclusion stated in paragraph 6 of this Declaration, I considered (a) the specification of the Tang '248 application, and (b) a number of published articles and patent documents that evidence gene expression monitoring techniques that were well-known before the November 5, 1998 filing date of the Tang '248 application. The published articles and patent documents I considered are:

(a) Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W., Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes, Proc. Natl. Acad. Sci. USA, 93, 10614-10619 (1996) (hereinafter "the Schena 1996 article") (copy annexed at Tab A);

(b) Schena, M., Shalon, D., Davis, R.W., Brown, P.O., Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray, Science, 270, 467-470 (1995) (hereinafter "the Schena 1995 article") (copy annexed at Tab B);

(c) Shalon and Brown PCT patent application WO 95/35505 titled "Method and Apparatus For Fabricating Microarrays Of Biological Samples," filed on June 16, 1995, and published on December 28, 1995 (hereinafter "the Shalon PCT application") (copy annexed at Tab C);

(d) Brown and Shalon U.S. Patent No. 5,807,522, corresponding to the Shalon PCT application, titled "Methods For Fabricating Microarrays Of Biological Samples," filed on June 7, 1995 and issued on September 15, 1998 (hereinafter "the Brown '522 patent") (copy annexed at Tab D);

(e) DeRisi, J., Penland, L., and Brown, P.O. (Group 1); Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (Group 2), Use of a cDNA microarray to analyse gene expression patterns in human cancer, Nat. Genet., 14(4), 457-460 (1996) (hereinafter "the DeRisi article") (copy annexed at Tab E);

(f) Shalon, D., Smith, S.J., and Brown, P.O., A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization, Genome Res., 6(7), 639-645 (1996) (hereinafter "the Shalon article") (copy annexed at Tab F);

(g) Heller, R.A., Schena, M., Chai A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E., and Davis R.W., Discovery and analysis of inflammatory disease-related genes using

cDNA microarrays, Proc. Natl. Acad. Sci. USA, 94, 2150-2155 (1997) (hereinafter “the Heller article”)(copy annexed at Tab G);

(h) Sambrook, J., Fritsch, E.F., Maniatis, T., Molecular Cloning, A Laboratory Manual, pages 7.37 and 7.38, Cold Spring Harbor Press (1989) (hereinafter “the Sambrook Manual”) (copy annexed at Tab H);

8. Many of the published articles and patent documents I considered (i.e., at least items (a)-(g) identified in paragraph 7) relate to work done at Stanford University in the early and mid-1990s with respect to the development of cDNA microarrays for use in gene expression monitoring applications under which Synteni became exclusively licensed. As I will discuss, a person skilled in the art who read the Tang ‘248 application on November 5, 1998 would have understood that application to disclose the SEQ ID NO:1-encoding polynucleotides to be useful for a number of gene expression monitoring applications, e.g., as a probe for the expression of that specific polynucleotide in cDNA microarrays of the type first developed at Stanford.

9. Turning more specifically to the Tang ‘248 specification, the SEQ ID NO:2 polynucleotide is shown at pp. 3-4 as one of 4 sequences under the heading “Sequence Listing.” The Tang ‘248 specification specifically teaches that the invention “provides an isolated and purified polynucleotide encoding the polypeptide comprising the amino acid sequence of SEQ ID NO:1” (Tang ‘248 application at p. 2). It further teaches that (a) the identity of the SEQ ID NO:2 polynucleotide was determined from a colon tumor tissue cDNA library (COLNTUT03) (Tang ‘248 application at pp. 11 and 33), (b) the SEQ ID NO:2 polynucleotide encodes for the human myosin heavy chain homolog (MHCH) shown as SEQ ID NO:1 (Tang ‘248 application at p. 11), and (c) northern analysis of SEQ ID NO:2 shows its expression predominantly in cDNA libraries associated with hematopoietic/immune system, gastrointestinal, musculoskeletal, and reproductive tissues, and in tissues associated with cancer (Specification at page 12, lines 4-9).

The Tang ‘248 application discusses a number of uses of the SEQ ID NO:1-encoding polynucleotides in addition to their use in gene expression monitoring applications. I have not fully evaluated these additional uses in connection with the preparation of this Declaration and do not express any views in this Declaration regarding whether or not the Tang ‘248 specification discloses

these additional uses to be substantial, specific and credible real-world utilities of the SEQ ID NO:1-encoding polynucleotides. Consequently, my discussion in this Declaration concerning the Tang '248 application focuses on the portions of the application that relate to the use of the SEQ ID NO:1-encoding polynucleotides in gene expression monitoring applications.

10. The Tang '248 application discloses that the polynucleotide sequences disclosed therein, including the SEQ ID NO:1-encoding polynucleotides, are useful as probes in microarrays. It further teaches that the microarrays can be used "to monitor the expression level of large numbers of genes simultaneously" for a number of purposes, including "to develop and monitor the activities of therapeutic agents" (Tang '248 application at p. 31, lines 35-36).

In the paragraph immediately following the Tang '248 teachings described in the preceding paragraph of this Declaration, the Tang '248 application teaches that microarrays can be prepared using the previously mentioned cDNA microarray technology developed at Stanford in the early to mid-1990s. In this connection, the Tang '248 application specifically cites to the Schena 1996 article identified in item (a) of paragraph 7 of this Declaration (Tang '248 application at p. 32; *supra*, paragraph 7).

The Schena 1996 article is one of a number of documents that were published prior to the November 5, 1998 filing date of the Tang '248 application that describes the use of the Stanford-developed cDNA technology in a wide range of gene expression monitoring applications, including monitoring and analyzing gene expression patterns in human cancer. In view of the Tang '248 application, the Schena 1996 article, and other related pre-November 1998 publications, persons skilled in the art on November 5, 1998 clearly would have understood the Tang '248 application to disclose the SEQ ID NO:1-encoding polynucleotides to be useful in cDNA microarrays for the development of new drugs and monitoring the activities of drugs for such purposes as evaluating their efficacy and toxicity, as explained more fully in paragraph 15 below.

With specific reference to toxicity evaluations, those of skill in the art who were working on drug development in November 1998 (and for many years prior to November 1998) without any doubt appreciated that the toxicity (or lack of toxicity) of any proposed drug they were working on was one of the most important criteria to be considered and evaluated in connection with

the development of the drug. They would have understood at that time that good drugs are not only potent, they are specific. This means that they have strong effects on a specific biological target and minimal effects on all other biological targets. Ascertaining that a candidate drug affects its intended target, and identification of undesirable secondary effects (i.e., toxic side effects), had been for many years among the main challenges in developing new drugs. The ability to determine which genes are positively affected by a given drug, coupled with the ability to quickly and at the earliest time possible in the drug development process identify drugs that are likely to be toxic because of their undesirable secondary effects, have enormous value in improving the efficiency of the drug discovery process, and are an important and essential part of the development of any new drug. Accordingly, the teachings in the Tang '248 application, in particular regarding use of the SEQ ID NO:1-encoding polynucleotides in differential gene expression analysis and in the development and the monitoring of the activities of drugs, clearly includes toxicity studies and persons skilled in the art who read the Tang '248 application on November 5, 1998 would have understood that to be so.

11. The Schena 1996 article was not the first publication that described the use of the cDNA microarray technique developed at Stanford to monitor quantitatively gene expression patterns. More than a year earlier (i.e., in October 1995), the Schena 1995 article, titled "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", was published (see Tabs A and B).

12. As previously discussed (*supra*, paragraphs 2 and 7), in the mid-1990s patent applications were filed in the names of Drs. Shalon and Brown that described the Stanford-developed cDNA microarray technology. The two patent documents (i.e., the Shalon PCT application and the Brown '522 patent) annexed to this Declaration at Tabs C and D evidence information that was available to the public regarding the Stanford-developed cDNA microarray technology before the November 5, 1998 filing date of the Tang '248 application.

The Shalon PCT patent application, which was published in December 1995, contains virtually the same (if not exactly the same) specification as the Brown '522 patent. Hence, the Brown '522 patent disclosure was, in effect, available to the public as of the December 1995 publication date

of the Shalon PCT application(see Tabs C and D). For the sake of convenience, I cite to and discuss the Brown '522 specification below on the understanding that the descriptions in that specification were published as of the December 28, 1995 publication date of the Shalon PCT application.

The Brown '522 patent discusses, in detail, the utility of the Stanford-developed cDNA microarrays in gene expression monitoring applications. For example, in the “Summary Of The Invention” section, the Brown '522 patent teaches (see Tab D, col. 4, line 52-col. 5, line 8):

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNAs from mRNAs isolated from two cells types, where the cDNAs from the first and second cell types are labeled with first and second different flourescent reporters.

A mixture of the labeled cDNAs from the two cell types is added to an array of polynucleotides representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNAs to complementary-sequence polynucleotides in the array. The array is then examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNAs derived from one of the first or second cell types give a distinct first and second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNAs derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes in the two cell types can then be determined by the observed fluorescence emission color of each spot.

The Brown '522 patent further teaches that the “[m]icroarrays of immobilized nucleic acid sequences prepared in accordance with the invention” can be used in “numerous” genetic applications, including “monitoring of gene expression” applications (see Tab D at col. 14, lines 36-42). The Brown '522 patent teaches (a) monitoring gene expression (i) in different tissue types, (ii) in different disease states, and (iii) in response to different drugs, and (b) that arrays disclosed therein may be used in toxicology studies (see Tab D at col. 15, lines 13-18 and 52-58 and col. 18, lines 25-30).

13. Also pertinent to my considerations underlying this Declaration is the DeRisi article, published in December 1996. The DeRisi article describes the use of the Stanford-developed cDNA microarray technology “to analyze gene expression patterns in human cancer” (see Tab E at, e.g., p. 457). The DeRisi article specifically indicates, consistent with what was apparent to persons skilled in the art in December 1996, that increasing the number of genes on the cDNA microarray permits a “more comprehensive survey of gene expression patterns,” thereby enhancing the ability of the cDNA microarray to provide “new and useful insights into human biology and a deeper understanding of the gene pathways involved in the pathogenesis of cancer and other diseases” (see Tab E at p. 458).

14. Other pre-November 1998 publications further evidence the utility of the cDNA microarrays first developed at Stanford in a wide range of gene expression monitoring applications (see, e.g., the Shalon and the Heller articles at Tabs F and G). By no later than the March 1997 publication of the Heller article, these publications showed that employees of Synteni (i.e., James Gilmore and myself) had used the cDNA microarrays in specific gene expression monitoring applications (see Tab G).

The Heller article states that the results reported therein “successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases” (Tab G at p. 2150). Among other things, the Heller article describes the investigation of “1000 human genes that were randomly selected from a peripheral human blood cell library” and “[t]heir differential and quantitative expression analysis in cells of the joint tissue. . . to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression” (see Tab G at pp. 2150 *et seq.*).

Much of the work reported on in the Heller article was done in 1996. That article, therefore, evidences how persons skilled in the art were readily able, well prior to November 5, 1998, to make and use cDNA microarrays to achieve highly useful results. For example, as reported in the Heller article, a cDNA microarray that was used in some of the highly successful work reported on therein was made from 1,000 genes randomly selected from a human blood cell library.

15. A person skilled in the art on November 5, 1998, who read the Tang '248 application, would understand that application to disclose the SEQ ID NO:1-encoding polynucleotides, for example, SEQ ID NO:2, to be highly useful as probes for the expression of that specific polynucleotide in cDNA microarrays of the type first developed at Stanford. For example, the specification of the Tang '248 application would have led a person skilled in the art in November 1998 who was using gene expression monitoring in connection with working on developing new drugs for the treatment of heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer to conclude that a cDNA microarray that contained the SEQ ID NO:1-encoding polynucleotides would be a highly useful tool and to request specifically that any cDNA microarray that was being used for such purposes contain the SEQ ID NO:1-encoding polynucleotides. Persons skilled in the art would appreciate that cDNA microarrays that contained the SEQ ID NO:1-encoding polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer for such purposes as evaluating their efficacy and toxicity.

I discuss in more detail in items (a)-(g) below a number of reasons why a person skilled in the art, who read the Tang '248 specification in November 1998, would have concluded based on that specification and the state of the art at that time, that the SEQ ID NO:1-encoding polynucleotides would be a highly useful tool for inclusion in cDNA microarrays for evaluating the efficacy and toxicity of proposed drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer, as well as for other evaluations:

(a) The Tang '248 application teaches the SEQ ID NO:1-encoding polynucleotides to be useful as probes in cDNA microarrays of the type first developed at Stanford. It also teaches that such cDNA microarrays are useful in a number of gene expression monitoring applications, including "developing and monitoring the activity of therapeutic agents [i.e., drugs]" (see paragraph 10, *supra*).

(b) By November 1998, the Stanford-developed cDNA microarray technology was a well known and widely accepted tool for use in a wide range of gene expression monitoring applications. This is evidenced, for example, by numerous publications describing the use of that

cDNA technology in gene expression monitoring applications and the fact that, for over a year, the technology had provided the basis for the operations of an up-and-running company (Synteni), with employees, that was created for the purpose of developing and commercially exploiting that technology (see paragraphs 2, 8 and 10-14, *supra*). The fact that Incyte agreed to purchase Synteni in late 1997 for an amount reported to be at least about \$80 million only serves to underscore the substantial practical and commercial significance, in 1997, of the cDNA microarray technology first developed at Stanford (see paragraph 2, *supra*).

(c) The pre-November 1998 publications regarding the cDNA microarray technology first developed at Stanford that I discuss in this Declaration repeatedly confirm that, consistent with the teachings in the Tang '248 application, cDNA microarrays are highly useful tools for conducting gene expression monitoring applications with respect to the development of drugs and the monitoring of their activity. Among other things, those pre-November 1998 publications confirmed that cDNA microarrays (i) were useful for monitoring gene expression responses to different drugs (see paragraph 12, *supra*), (ii) were useful in analyzing gene expression patterns in human cancer, with increasing the number of genes on the cDNA microarray enhancing the ability of the cDNA microarray to provide useful information (see paragraph 13, *supra*), and (iii) were a valuable tool for use as part of a "general approach for dissecting human diseases" and for "analyz[ing] complex diseases by their pattern of gene expression" (see paragraph 14, *supra*).

(d) Based on my own extensive work for a company whose business was the development and commercial exploitation of cDNA microarray technology for more than two years prior to the November 1998 filing date of the Tang '248 application, I have first-hand knowledge concerning the state of the art with respect to making and using cDNA microarrays as of November 5, 1998 (see paragraphs 2 and 14, *supra*). Persons skilled in the art as of that date would have (a) concluded that the Tang '248 application disclosed cDNA microarrays containing the SEQ ID NO:1-encoding polynucleotides to be useful, and (b) readily been able to make and use such microarrays with useful results.

(e) The Tang '248 specification contains a number of teachings that would lead persons skilled in the art on November 5, 1998 to conclude that a cDNA microarray that contained the SEQ ID NO:1-encoding polynucleotides would be a more useful tool for gene expression monitoring

applications relating to drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer than a cDNA microarray that did not contain the SEQ ID NO:1-encoding polynucleotides. Among other things, the Tang '248 specification teaches that the identity of the SEQ ID NO:2 polynucleotide was determined from a colon tumor tissue cDNA library (COLNTUT03) (Tang '248 application at pp. 11 and 33). Moreover, northern analysis of SEQ ID NO:2 shows its expression predominantly in cDNA libraries associated with hematopoietic/immune system, gastrointestinal, musculoskeletal, and reproductive tissues, and in tissues associated with cancer (Specification at page 12, lines 4-9). (See paragraph 9, *supra*).

Moreover, the Tang '248 specification teaches that the MHCH protein having the amino acid sequence of SEQ ID NO:1 shares homology with known functional proteins. MHCH is a member of the human receptor protein family. In particular, SEQ ID NO:1 shares homology with the sequence of *C. elegans* myosin (g1279777) and *H. annuus* unconventional myosin (g2444174). (Tang '248 application, at p. 11).

(f) Persons skilled in the art on November 5, 1998 would have appreciated (i) that the gene expression monitoring results obtained using a cDNA microarray containing a probe to a sequence selected from the group consisting of SEQ ID NO:1-encoding polynucleotides would vary, depending on the particular drug being evaluated, and (ii) that such varying results would occur both with respect to the results obtained from the probe described in (i) and from the cDNA microarray as a whole (including all its other individual probes). These kinds of varying results, depending on the identity of the drug being tested, in no way detracts from my conclusion that persons skilled in the art on November 5, 1998, having read the Tang '248 specification, would specifically request that any cDNA microarray that was being used for conducting gene expression monitoring studies on drugs for treating heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer (*e.g.*, a toxicology study or any efficacy study of the type that typically takes place in connection with the development of a drug) contain any one of the SEQ ID NO:1-encoding polynucleotides as a probe. Persons skilled in the art on November 5, 1998 would have wanted their cDNA microarray to have a probe as described in (i) because a microarray that contained such a probe (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies

using cDNA microarrays that persons skilled in the art have been doing since well prior to November 5, 1998.

The foregoing is not intended to be an all-inclusive explanation of all my reasons for reaching the conclusions stated in this paragraph 15, and in paragraph 6, *supra*. In my view, however, it provides more than sufficient reasons to justify my conclusions stated in paragraph 6 of this Declaration regarding the Tang '248 application disclosing to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the SEQ ID NO:1-encoding polynucleotides.

16. Also pertinent to my considerations underlying this Declaration is the fact that the Tang '248 disclosure regarding the uses of the SEQ ID NO:2 polynucleotide for gene expression monitoring applications is not limited to the use of that polynucleotide as a probe in microarrays. For one thing, the Tang '248 disclosure regarding the hybridization technique used in gene expression monitoring applications is broad (Tang '248 application at, e.g., p. 3, lines 4-9).

In addition, the Tang '248 specification repeatedly teaches that the polynucleotides described therein (including the polynucleotide of SEQ ID NO:2) may desirably be used as probes in any of a number of long established "standard" non-microarray techniques, such as Northern analysis, for conducting gene expression monitoring studies. See, e.g.:

(a) Tang '248 application at p. 7, lines 11-13 ("[N]orthern analysis is indicative of the presence of nucleic acids encoding MHCH in a sample, and thereby correlates with expression of the transcript from the polynucleotide encoding MHCH");

(b) Tang '248 application at p. 30, lines 24-27 ("The polynucleotide sequences encoding MHCH may be used in Southern or northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; in dipstick, pin, and multiformat ELISA-like assays; and in microarrays utilizing fluids or tissues from patients to detect altered MHCH expression. Such qualitative or quantitative methods are well known in the art");

(c) Tang '248 application at p. 31, lines 1-10 ("In order to provide a basis for the diagnosis of a disorder associated with expression of MHCH, a normal or standard profile for expression is established. This may be accomplished by combining body fluids or cell extracts taken

from normal subjects, either animal or human, with a sequence, or a fragment thereof, encoding MHCH, under conditions suitable for hybridization or amplification. Standard hybridization may be quantified by comparing the values obtained from normal subjects with values from an experiment in which a known amount of a substantially purified polynucleotide is used. Standard values obtained in this manner may be compared with values obtained from samples from patients who are symptomatic for a disorder. Deviation from standard values is used to establish the presence of a disorder”); and

(d) Tang ‘248 application at p. 35, lines 14-17 (“Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound. (See, e.g., Sambrook, supra, ch. 7; Ausubel, 1995, supra, ch. 4 and 16.)”).

The “Sambrook et al.” reference cited in item (d) immediately above is a reference that was well known to persons skilled in the art in November 1998. A copy of pages from that reference manual, which was published in 1989, is annexed to this Declaration at Tab H. The attached pages from the Sambrook manual provide an overview of northern analysis and other membrane-based technologies for conducting gene expression monitoring studies that were known and used by persons skilled in the art for many years prior to the November 5, 1998 filing date of the Tang ‘248 application.

A person skilled in the art on November 5, 1998, who read the Tang ‘248 specification, would have routinely and readily appreciated that the SEQ ID NO:1-encoding polynucleotides disclosed therein would be useful as a probe to conduct gene expression monitoring analyses using northern analysis or any of the other traditional membrane-based gene expression monitoring techniques that were known and in common use many years prior to the filing of the Tang ‘248 application. For example, a person skilled in the art in November 1998 would have routinely and readily appreciated that the SEQ ID NO:1-encoding polynucleotides would be a useful tool in conducting gene expression analyses, using the northern analysis technique, in furtherance of (a) the development of drugs for the treatment of heart and skeletal muscle disorders, developmental disorders, and cell proliferative disorders, including cancer, and (b) analyses of the efficacy and toxicity of such drugs.

PCTWORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶: G01N 33/543, 33/68	A1	(11) International Publication Number: WO 95/35505 (43) International Publication Date: 28 December 1995 (28.12.95)
(21) International Application Number: PCT/US95/07659 (22) International Filing Date: 16 June 1995 (16.06.95) (30) Priority Data: 08/261,388 17 June 1994 (17.06.94) US 08/477,809 7 June 1995 (07.06.95) US (71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Stanford, CA 94305 (US). (72) Inventors: SHALON, Tidhar, Dari; 364 Fletcher Drive, Atherton, CA 94027 (US). BROWN, Patrick, O.; 76 Peter Coutts Circle, Stanford, CA 94305 (US). (74) Agent: DEHLINGER, Peter, J.; Dehlinger & Associates, P.O. Box 60850, Palo Alto, CA 94306-1546 (US).		(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i>
(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES (57) Abstract A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

**METHOD AND APPARATUS FOR FABRICATING
MICROARRAYS OF BIOLOGICAL SAMPLES**

Field of the Invention

5 This invention relates to a method and apparatus
for fabricating microarrays of biological samples for
large scale screening assays, such as arrays of DNA
samples to be used in DNA hybridization assays for
genetic research and diagnostic applications.

10

References

Abouzied, et al., *Journal of AOAC International*
77(2):495-500 (1994).

Bohlander, et al., *Genomics* 13:1322-1324 (1992).

15 Drmanac, et al., *Science* 260:1649-1652 (1993).

Fodor, et al., *Science* 251:767-773 (1991).

Khrapko, et al., *DNA Sequence* 1:375-388 (1991).

Kuriyama, et al., AN ISFET BIOSENSOR, APPLIED BIOSENSORS
(Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).

20 Lehrach, et al., HYBRIDIZATION FINGERPRINTING IN GENOME
MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and
Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81
(1990).

25 Maniatis, et al., MOLECULAR CLONING, A LABORATORY
MANUAL, Cold Spring Harbor Press (1989).

Nelson, et al., *Nature Genetics* 4:11-18 (1993).

Pirrung, et al., U.S. Patent No. 5,143,854 (1992).

Riles, et al., *Genetics* 134:81-150 (1993).

Schena, M. et al., *Proc. Nat. Acad. Sci. USA*
89:3894-3898 (1992).

5 Southern, et al., *Genomics* 13:1008-1017 (1992).

Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as
10 arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells
15 to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation.
20 This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

25 A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array
30 includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 x 22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, et al. (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatis, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a
5 nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the
10 immobilized antibody and an antigen is detected using a standard ELISA colorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose
15 somewhat hydrophobic using a line drawn with PAP Pen (Research Products International). However Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the
20 hydrophobic barrier during extended high temperature incubations or in the presence of detergents which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of
25 hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego CA) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since
30 reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate.
35 These plates are capable of containing different

reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many colorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barriers elements are shallow and do not interfere with the detection step thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

Summary of the Invention

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous

solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

5 The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are
10 repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is are applied to a selected position on each of a plurality of solid supports at each repeat cycle.

15 The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new
20 reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected,
25 analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes positioning
30 structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a

selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm². Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on

the backing, and a grid formed on the film. The grid is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impervious cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impervious grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA'S from the first and second cells are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNA's from the two cell types is added to an array of polynucleotides

representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array. The array is then
5 examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color,
10 respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes
15 in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read
20 in conjunction with the accompanying figures.

Brief Description of the Drawings

Fig. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head
25 constructed for use in one embodiment of the invention;

Figs. 2A-2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from Fig. 1, in accordance with one embodiment of the method of the invention;

30 Fig. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

Fig. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance
35 with the invention.

Fig. 5 shows a fluorescent image of an actual 20 x 20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-l-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm x 1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

Fig. 7 shows the translation of the hybridization image of Fig. 6 into a karyotype of the yeast genome, where the elements of Fig.-6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm x 0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type *Arabidopsis* plant labeled with a green fluorophore and total cDNA from a transgenic *Arabidopsis* plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic *Arabidopsis* plant;

Fig. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

Fig. 10 shows an enlarged plan view of one of the cells in the substrate in Fig. 9, showing an array of polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the substrate in Fig. 9, taken along a section line in that figure; and

Fig. 12 is a scanned image of a 3 cm x 3 cm nitrocellulose solid support containing four identical

arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

5

Detailed Description of the Invention

I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

10 "Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; an effector molecule in an effector/receptor binding pair;
15 or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a ligand/anti-ligand binding pair. The antiligand may be the other of the nucleic acid strands in a
20 complementary, hybridized nucleic acid duplex binding pair; the receptor molecule in an effector/receptor binding pair; or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

25 "Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

30 "Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a
35 linear or two-dimensional array of preferably discrete

regions, each having a finite area, formed on the surface of a solid support.

5 A "microarray" is an array of regions having a density of discrete regions of at least about $100/\text{cm}^2$, and preferably at least about $1000/\text{cm}^2$. The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μm , and are separated from other regions in the array by about the same distance.

10 A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

15 A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the
20 biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-
25 concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides,
30 and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of

differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

5 This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

10 Fig. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below.

15 The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two

20 elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in Fig. 2A. The advantages of the open channel

25 construction of the dispenser are discussed below.

 With continued reference to Fig. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in

30 the dispenser on a support, as will be described below with reference to Figs. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring

35 bias, to a normal, raised position, as shown. The

dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

Figs. 2A-2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, i.e., wettable surface, such as a surface having native, bound or covalently attached charged groups. On such surface described below is a glass surface having an absorbed layer of a polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, i.e., one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved

to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the
5 dispenser tip in its raised position, as seen in Fig. 2A, where the tip is typically at least several 1-5 mm above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move
10 rapidly toward and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip
15 channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in Fig. 2B.

20 Fig. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size,
25 i.e., volume, the tendency of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends
30 toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, e.g., the positions illustrated in Figs. 2B and 2C.

The desired deposition volume, i.e., bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, i.e., the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in Fig. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μm .

Table 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

10 At a given tip size, bead volume can be reduced in
a controlled fashion by increasing surface
hydrophobicity, reducing time of contact of the tip
with the surface, increasing rate of movement of the
tip away from the surface, and/or increasing the
15 viscosity of the medium. Once these parameters are
fixed, a selected deposition volume in the desired pl
to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location
on a support, the tip is typically moved to a
20 corresponding position on a second support, a droplet
is deposited at that position, and this process is
repeated until a liquid droplet of the reagent has been
deposited at a selected position on each of a plurality
of supports.

25 The tip is then washed to remove the reagent
liquid, filled with another reagent liquid and this
reagent is now deposited at each another array position
on each of the supports. In one embodiment, the tip is
washed and refilled by the steps of (i) dipping the
30 capillary channel of the device in a wash solution,
(ii) removing wash solution drawn into the capillary
channel, and (iii) dipping the capillary channel into
the new reagent solution.

From the foregoing, it will be appreciated that
35 the tweezers-like, open-capillary dispenser tip

provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in Fig. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20-200 μm . The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20-400 μm . Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/cm². After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in Fig. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to Fig. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the

figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5-25 μm . In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more, for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5-250 μm , and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of reagent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuation of the dispenser is

then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 μ l and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

35 IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a
5 microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

10

A. Multi-Cell Substrate

Fig. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8 x 12 rectangular array 112 of cells, such as
15 cells 114, 116, formed on the substrate surface. With reference to Fig. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such
20 regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in Fig. 3.

The 96-cell array shown in Fig. 9 has typically array dimensions between about 12 and 244 mm in width
25 and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, i.e., between about 1 and 20 in width and 1 and 50 mm in length.

30 The construction of substrate is shown cross-sectionally in Fig. 11, which is an enlarged sectional view taken along view line 124 in Fig. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed
35 on the surface of the backing is a water-permeable film

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm . The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In Fig. 11, defined volume samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the

microarrays in the well are identical arrays of distinct biopolymers, e.g., different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by
5 depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

10 In a preferred embodiment, each microarray contains about 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . Also in a preferred embodiment, the biopolymers in each microarray region are present in a
15 defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method
20 described in Section II.

Also in a preferred embodiments, the biopolymers are polynucleotides having lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by schemes
25 involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded
30 onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard colorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

Fig. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between

negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-l-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, MO).

5 To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide
10 surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is
15 illustrated in Examples 1 and 2.

To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization
20 conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

25 In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . In the embodiment shown in Fig. 5, the microarray contains 400 regions in an area of about 16 mm^2 , or 2.5×10^3 regions/ cm^2 . Also in a preferred
30 embodiment, the polynucleotides in the each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-

density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

5 Also in a preferred embodiments, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various in situ synthesis schemes.

10

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous
15 genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment
20 is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described
25 by Nelson, et al. (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous
30 to the immobilized clones on the array. For example, Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene or
35 genes can be probed with a labeled mixture of a

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm x 2.2 cm microarrays fabricated on a single 12 cm x 18 cm

sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatis, et al., 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug cogener preparations or

chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

5 The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass
10 screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

15 The following examples illustrate, but in no way are intended to limit, the present invention.

Example 1

20 Genomic-Complexity Hybridization to Micro
DNA Arrays Representing the Yeast
Saccharomyces cerevisiae Genome with
Two-Color Fluorescent Detection

 The array elements were randomly amplified PCR
25 (Bohlander, et al., 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA templates (Riles, et al., 1993). The PCR was performed directly on the lambda phage lysates resulting in an amplification of both the 35 kb lambda vector and the
30 5-15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250-1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ) and concentrated by evaporation to dryness at room
35 temperature overnight. Each of the 864 amplified

lambda clones was rehydrated in 15 μ l of 3 \times SSC in preparation for spotting onto the glass.

5 The micro arrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine (Sigma). The automated apparatus described in Section IV loaded 1 μ l of the concentrated lambda clone PCR product in 3 \times SSC directly from 96 well storage plates into the open capillary printing element and deposited
10 -5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a
15 dry 80° vacuum oven for 2 hours, rinsed to remove unabsorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90°
20 for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, CA). The six largest chromosomes were isolated in one gel
25 slice and the smallest 10 chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, CA). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following
30 amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, IL) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, MA) for the pool containing the
35 six largest chromosomes, and with a fluorescein

conjugated nucleotide analog (BMB) for the pool containing smallest ten chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, MA).

5 Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 μ l of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 μ l of concentrated hybridization solution (5 \times SSC and 0.1% SDS) was added and all 10 μ l
10 transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1 \times SSC and 0.1%SDS for 5 minutes, cover slipped and scanned.

15 A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8 \times 1.8 cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical
20 crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype
25 of the yeast genome.

Figure 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the largest six yeast
30 chromosomes. A green signal indicates that the lambda clone insert comes from one of the smallest ten yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the
35 hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, et al.) allowing for the automatic generation of the color karyotype shown in Figure 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the largest six chromosomes are mainly red while the smallest ten chromosomes are mainly green matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into a position on the physical map of yeast.

25

Example 2

Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

24 clones containing cDNA inserts from the plant *Arabidopsis* were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3 x SSC. The cDNA clones were spotted on poly-l-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone

35

representing a transcription factor HAT 4, which had previously been used to create a transgenic line of the plant *Arabidopsis*, in which this gene is present at ten times the level found in wild-type *Arabidopsis* (Schena, et al., 1992).

Total poly-A mRNA from wild type *Arabidopsis* was isolated using standard methods (Maniatis, et al., 1989) and reverse transcribed into total cDNA, using fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of *Arabidopsis* where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lissamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. Fig. 8 show the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic *Arabidopsis* appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of *Arabidopsis* but not detectably expressed in wild type *Arabidopsis*, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic *Arabidopsis* and the relative lack of expression of the

transcription factor in the green-labeled wild type *Arabidopsis*.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

Example 3

Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barrier elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10 × SSC and allowed to dry. As shown in Fig. 12, 192 M13 clones each with a different yeast inserts were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of

each array are positive controls for the colorimetric detection step.

The oligonucleotides were labeled with fluorescein which was detected using an anti-fluorescein antibody
5 conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II)
10 attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable
15 dimples on the nitrocellulose which can be used to automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be
20 clear that various changes and modification may be made without departing from the invention.

IT IS CLAIMED:

1. A method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent, said method comprising,
 - (a) loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,
 - (b) tapping the tip of the dispensing device against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume of solution on the surface, and
 - (c) repeating steps (a) and (b) until said array is formed.
2. The method of claim 1, wherein said tapping is carried out with an impulse effective to deposit a selected volume in the volume range between 0.01 to 100 nl.
3. The method of claim 1, wherein said channel is formed by a pair of spaced-apart tapered elements.
4. The method of claim 1, for forming a plurality of such arrays, wherein step (b) is applied to a selected position on each of a plurality of solid supports at each repeat cycle proceeding step (c).

5 5. The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

10 6. Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising

15 (a) a holder for holding, at known positions, a plurality of planar supports,

 (b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

20 (c) positioning means for positioning the dispensing device at a selected array position with respect to a support in said holder,

25 (d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and

30 (e) control means for controlling said positioning and dispensing means.

7. The apparatus of claim 6, wherein said dispensing means is effective to move said dispensing device against a support with an impulse effective to deposit a selected volume in the volume range between
5 0.01 to 100 nl.

8. The apparatus of claim 6, wherein said channel is formed by a pair of spaced-apart tapered elements.

10 9. The apparatus of claim 6, wherein the control means operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and
15 (iii) dispense the reagent at a defined array position on each of the supports on said holder.

10. The apparatus of claim 6, wherein the control device further operates, at the end of a dispensing
20 cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing
25 device with a fresh selected reagent.

11. The apparatus of claim 6, wherein said device is one of a plurality of such devices which are carried on the arm for dispensing different analyte assay
30 reagents at selected spaced array positions.

12. A substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers per 1 cm^2 surface area, each

distinct biopolymer sample (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about 0.1 femtomole and 100 nanomoles.

13. The substrate of claim 12, wherein said surface is glass slide coated with polylysine, and said biopolymers are polynucleotides.

14. The substrate of claim 12, wherein said substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where said grid (i) is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and (ii) partitions the film into a plurality of water-impervious cells, where each cell contains such a biopolymer array.

15. A substrate with a surface array of sample-receiving cells, comprising
a water-impermeable backing,
a water-permeable film formed on the backing, and
a grid formed on the film, said grid being composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film.

16. The substrate of claim 15, wherein the cells of the array each contain an array of biopolymers.

17. A substrate for use in detecting binding of labeled biopolymers to one or more of a plurality distinct polynucleotides, comprising

a non-porous, glass substrate,
a coating of a cationic polymer on said substrate,
and

an array of distinct polynucleotides to said
5 coating, where each biopolymer is disposed at a
separate, defined position in a surface array of
biopolymers.

18. A method of detecting differential expression
10 of each of a plurality of genes in a first cell type
with respect to expression of the same genes in a
second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's
isolated from the two cells types, where the cDNA's
15 from the first and second cells are labeled with first
and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the
two cell types to an array of polynucleotides
representing a plurality of known genes derived from
20 the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array; and

examining the array by fluorescence under
fluorescence excitation conditions in which (i)
25 polynucleotides in the array that are hybridized
predominantly to cDNA's derived from one of the first
and second cell types give a distinct first or second
fluorescence emission color, respectively, and (ii)
polynucleotides in the array that are hybridized to
30 substantially equal numbers of cDNA's derived from the
first and second cell types give a distinct combined
fluorescence emission color, respectively,

wherein the relative expression of known genes in
the two cell types can be determined by the observed
35 fluorescence emission color of each spot.

19. The method of claim 18, wherein the array of polynucleotides is formed on a substrate with a surface having an array of at least 10^2 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm^2 , each distinct biopolymer (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about .1 femtomole and 100 nmoles.

10

20. The method of claim 19, wherein said surface is a glass slide coated with polylysine, and said biopolymers are polynucleotides non-covalently bound to said polylysine.

15

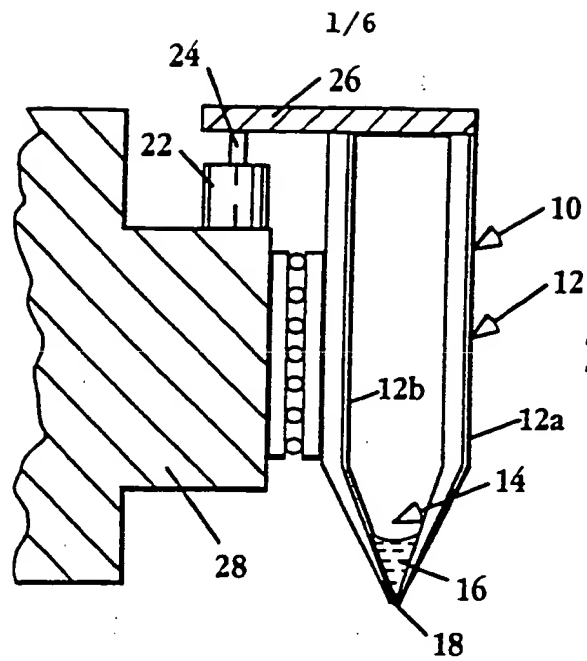


Fig. 1

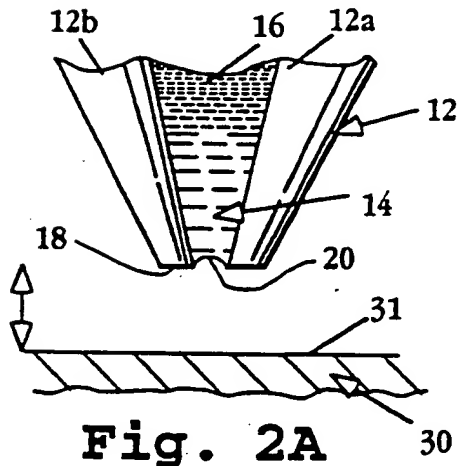


Fig. 2A

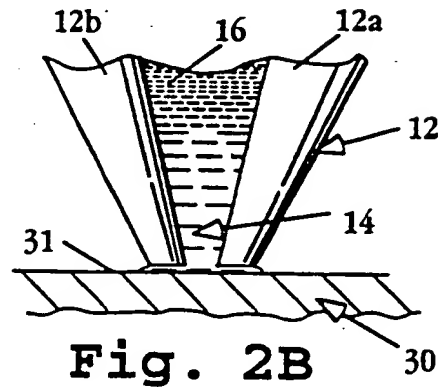


Fig. 2B

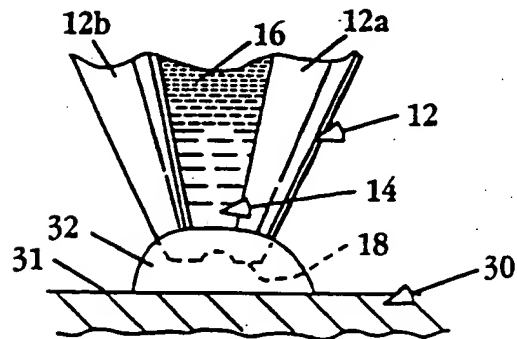
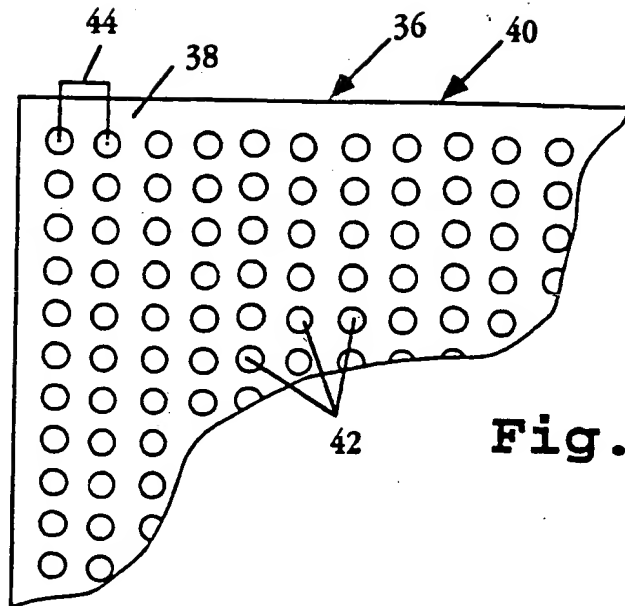
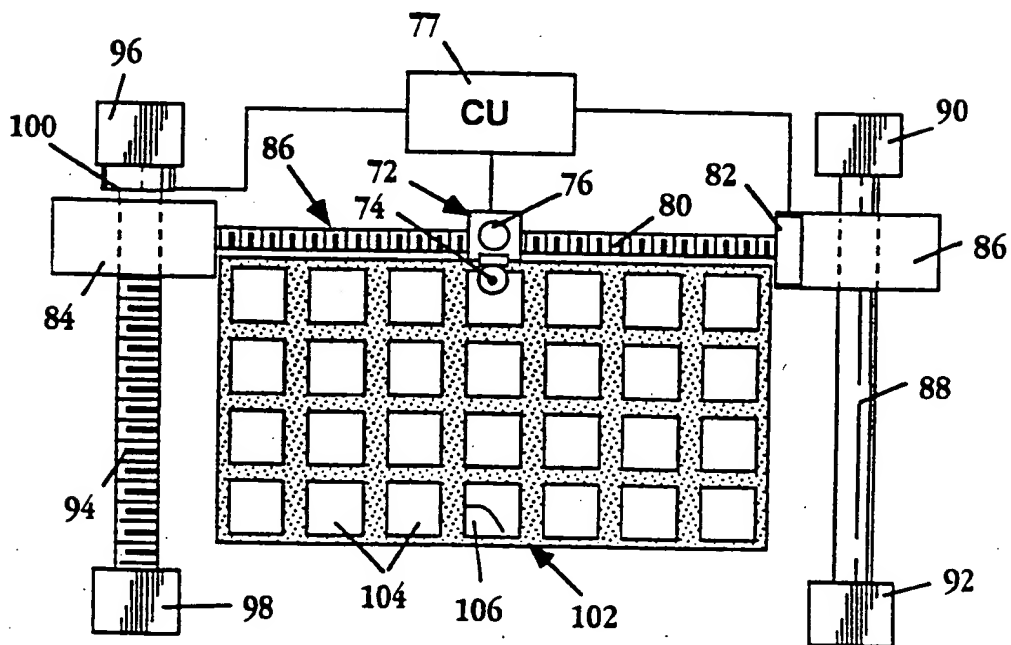


Fig. 2C

2/6

**Fig. 3****Fig. 4**

3/6

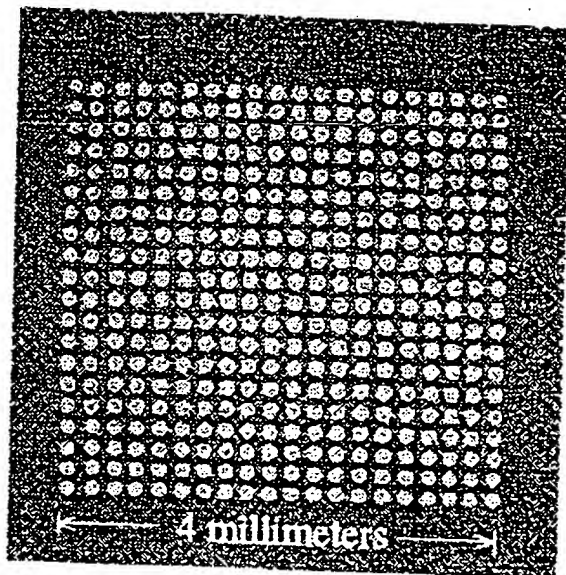


Fig. 5

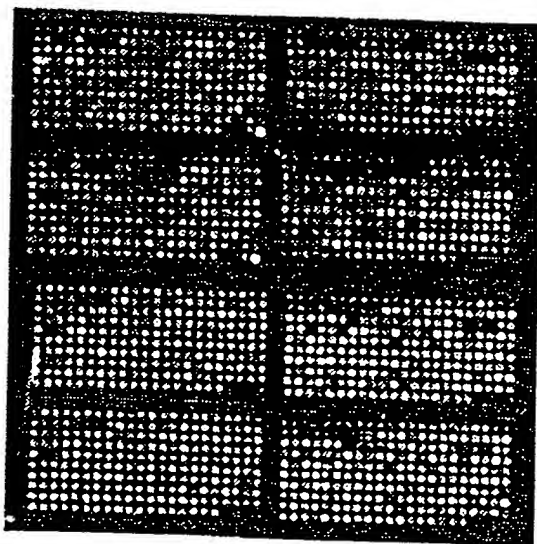


Fig. 6

4/6

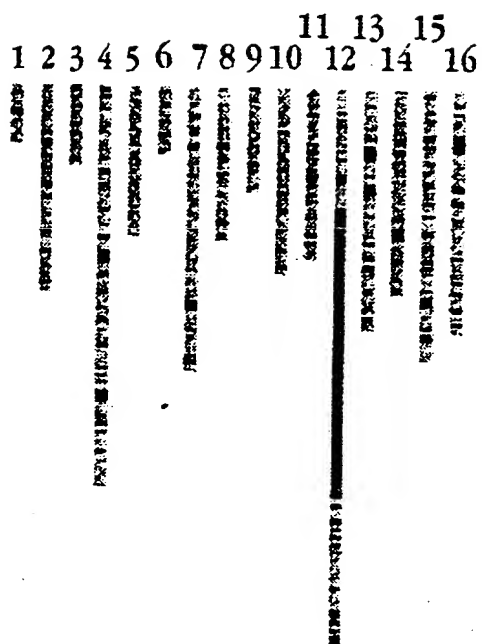


Fig. 7

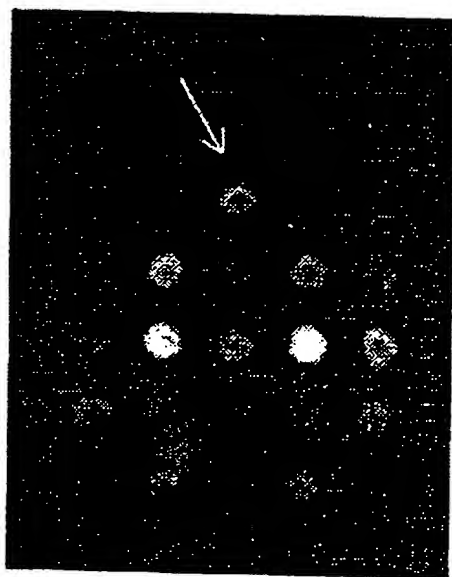


Fig. 8

5/6

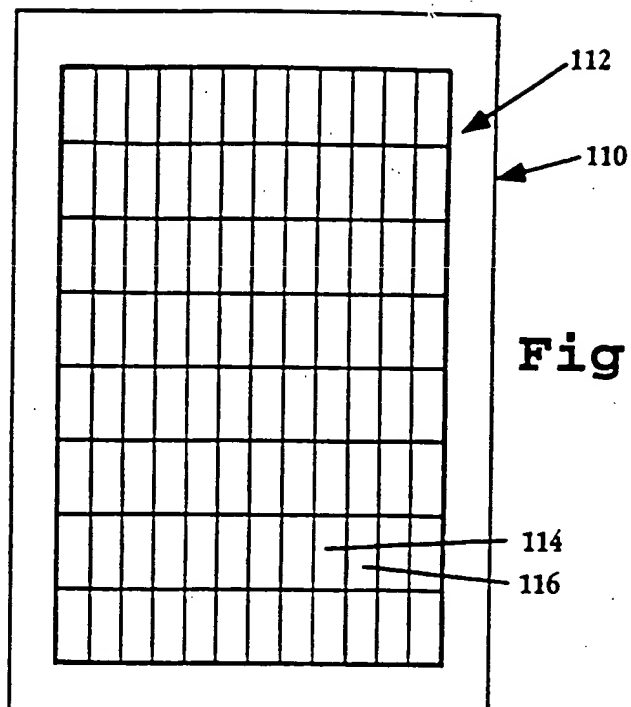


Fig. 9

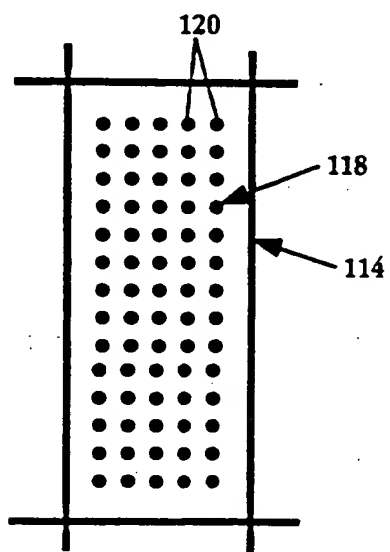


Fig. 10

6/6

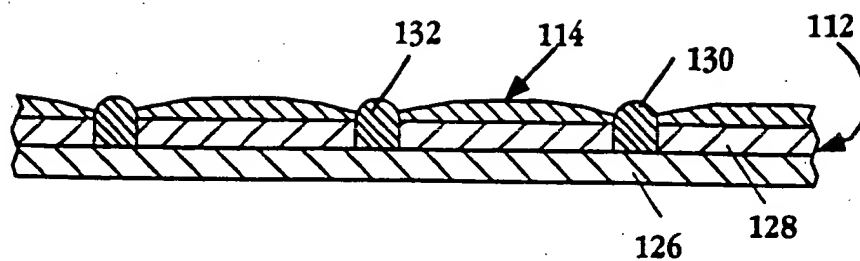


Fig. 11

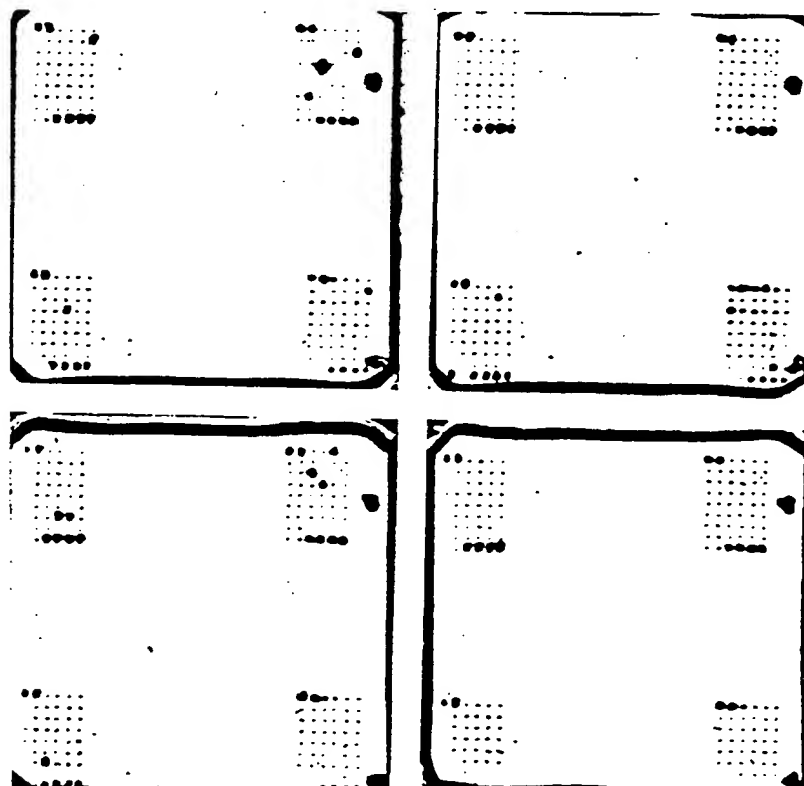


Fig. 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/07659

A. CLASSIFICATION OF SUBJECT MATTER		
IPC(6) : G01N 33/543, 33/68 US CL : 435/6; 436/518 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) U.S. : 422/57; 435/4.6.973; 436/518.524.527.531.805.809		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document	1-17
A	US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document.	6-11
A	US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document.	12-17
A	US, A, 5,100,777 (CHANG) 31 March 1992, see entire document.	12-17
A	US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document.	12-17

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

<p>* Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"Z" document member of the same patent family</p>
--	---

Date of the actual completion of the international search 15 SEPTEMBER 1995	Date of mailing of the international search report 06 OCT 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer CHRISTOPHER CHIN Telephone No. (703) 308-0196



US005807522A

United States Patent [19]

Brown et al.

[11] Patent Number: 5,807,522

[45] Date of Patent: Sep. 15, 1998

[54] METHODS FOR FABRICATING
MICROARRAYS OF BIOLOGICAL SAMPLES[75] Inventors: Patrick O. Brown, Stanford; Tldhar
Dart Sharon, Albeton, both of Calif.[73] Assignee: The Board of Trustees of the Leland
Stanford Junior University, Stanford,
Calif.

[21] Appl. No.: 477,809

[22] Filed: Jun. 7, 1995

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 261,388, Jun. 17, 1994,
abandoned.[51] Int. Cl.⁶ C12M 1/34; C12M 1/40[52] U.S. Cl. 422/50; 422/52; 422/55;
422/56; 422/57; 422/68.1; 422/69; 422/82.05;
422/82.06; 422/82.07; 422/82.08; 435/6;
435/7.1; 436/501; 530/300; 530/333; 530/334;
530/350; 536/25.3[58] Field of Search 435/6, 7.1, 172.3;
536/23.1, 24.31, 25.3; 935/78, 3, 19, 80;
436/501, 813; 422/50, 52, 55, 56, 57, 68.1,
69, 82.05, 82.06-82.08; 530/300, 333, 334,
350

[56] References Cited

U.S. PATENT DOCUMENTS

3,730,844	5/1973	Gilham et al.	435/6
4,071,315	1/1978	Chateau	436/518
4,486,539	12/1984	Ranki et al.	436/504
4,556,643	12/1985	Pearl et al.	435/5
4,563,419	1/1986	Ranki et al.	435/6
4,591,570	5/1986	Chang	436/518
4,670,380	6/1987	Dattagupta	435/6
4,677,054	6/1987	White et al.	435/6
4,683,195	7/1987	Mullis et al.	435/6
4,683,202	7/1987	Mullis	435/91.2
4,716,106	12/1987	Chiswell	435/6

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

721016A2	7/1996	European Pat. Off.
WO 90/03382	4/1990	WIPO
WO 92/10588	6/1992	WIPO
WO 93/22680	11/1993	WIPO
WO 95/00530	1/1995	WIPO
WO 95/15970	6/1995	WIPO
WO 95/21944	8/1995	WIPO
WO 95/25116	9/1995	WIPO
WO 96/17958	6/1996	WIPO

OTHER PUBLICATIONS

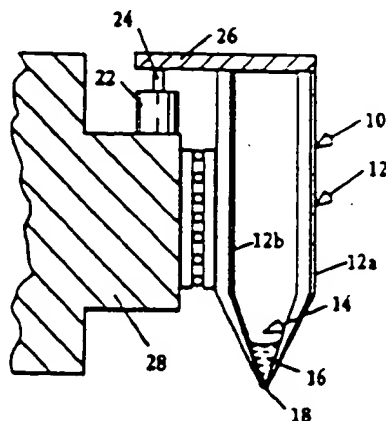
Billings et al., "New Techniques for Physical Mapping of the
Human Genome," *FASEB*, 5:28-34 (1991).Chee, et al., "Accessing Genetic Information with High-
Density DNA Arrays", *Science*, 274:610-614 (1996).Drmanac et al., "DNA Sequence Determination by Hybrid-
ization: A Strategy for Efficient Large-Scale Sequencing,
" *Science*, 260:1649-1652 (1993).Drmanac et al., "Laboratory Methods: Reliable Hybridiza-
tion of Oligonucleotides as Short as Six Nucleotides," *DNA
and Cell Biology*, 9:527-534 (1990).Drmanac et al., "Sequencing by Hybridization: Towards an
Automated Sequencing of One Million M13 Clones Arrayed
on Membranes," *Electrophoresis*, 13:566-573 (1992).Elkins, et al., "Multianalyte Immunoassay: The Immunologi-
cal 'Compact Disk' of the Future", *J. Clinical Immunoassay*,
13(4):169-181 (1990).

Primary Examiner—Ardin H. Marschel

Attorney, Agent, or Firm—Arnold White & Durkee

[57] ABSTRACT

A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each selected array position; by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.

7 Claims, 6 Drawing Sheets
(2 of 6 Drawing(s) Filed in Color)

U.S. PATENT DOCUMENTS

4,731,325	3/1988	Pajva et al.	435/6	5,328,824	7/1994	Ward et al.	435/6
4,755,458	7/1988	Rabbani et al.	435/5	5,338,686	8/1994	Deeg et al.	436/180
4,767,700	8/1988	Wallace	435/6	5,348,855	9/1994	Dattagupta et al.	435/6
4,868,104	9/1989	Kura et al.	435/6	5,389,512	2/1995	Soisakky et al.	435/5
4,868,105	9/1989	Urdes et al.	435/6	5,412,087	5/1995	McGill et al.	536/24.3
4,921,805	5/1990	Gebevech et al.	435/270	5,434,049	7/1995	Okano et al.	435/6
4,981,783	1/1991	Augenlicht	435/6	5,445,934	8/1995	Fodor et al.	435/6
5,013,669	5/1991	Peters, Jr. et al.	436/518	5,472,842	12/1995	Stokke et al.	435/6
5,028,545	7/1991	Soini	436/501	5,474,796	12/1995	Brennan	427/2.13
5,064,754	11/1991	Mills	435/6	5,474,895	12/1995	Ishii et al.	435/6
5,091,652	2/1992	Mathies et al.	250/458.1	5,510,270	4/1996	Fodor et al.	436/518
5,100,777	3/1992	Chang	435/7.24	5,512,430	4/1996	Goag	435/5
5,143,854	9/1992	Pfarrung et al.	436/518	5,514,543	5/1996	Grossman et al.	435/6
5,185,243	2/1993	Ullman et al.	435/6	5,514,785	5/1996	Van Nieu et al.	536/22.1
5,188,963	2/1993	Stapleton	435/288.3	5,516,641	5/1996	Ullman et al.	435/6
5,200,051	4/1993	Cozzette et al.	204/407	5,518,883	5/1996	Soiai	435/6
5,200,312	4/1993	Oprandy	435/5	5,545,531	8/1996	Rava et al.	435/6
5,202,231	4/1993	Dronamur et al.	435/6	5,556,748	9/1996	Douglas	435/6
5,204,268	4/1993	Matsu moto	436/44	5,556,752	9/1996	Lockhart et al.	435/6
5,242,974	9/1993	Holmes	525/54.11	5,563,060	10/1996	Hozier	435/240.23
5,252,296	10/1993	Zuckerman et al.	422/116	5,578,832	11/1996	Trubon et al.	250/458.1
5,252,743	10/1993	Barrett et al.	548/303.7	5,605,662	2/1997	Heller et al.	422/68.1

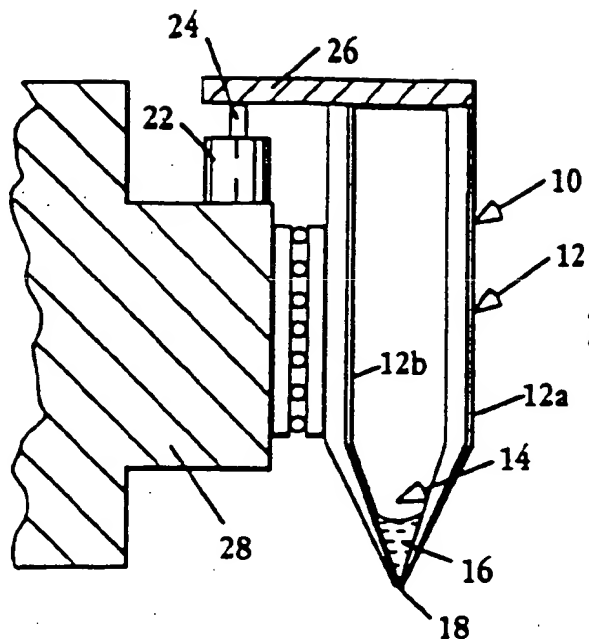


Fig. 1

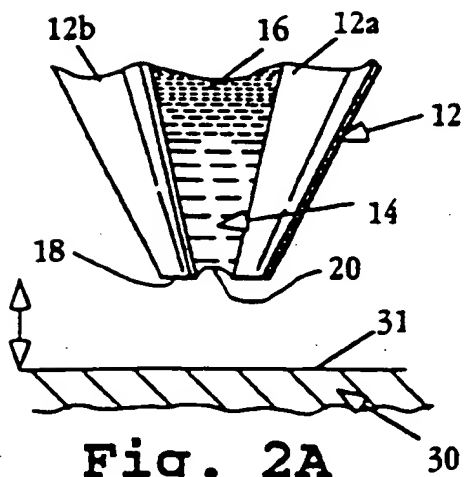


Fig. 2A

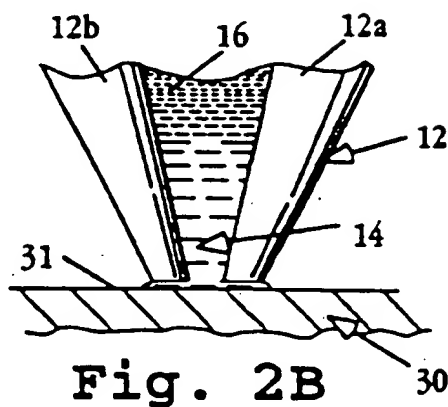


Fig. 2B

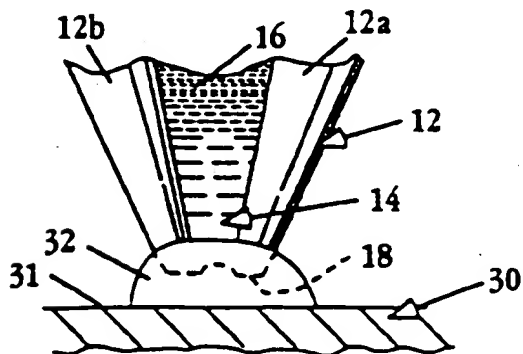


Fig. 2C

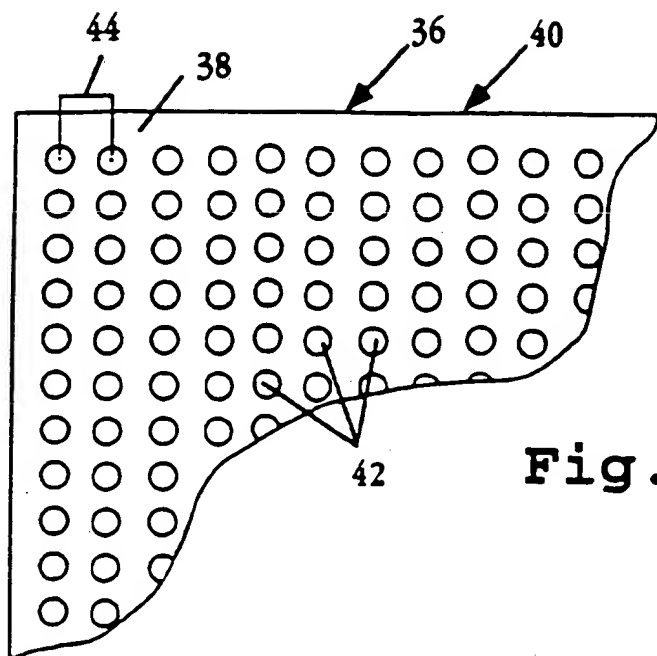


Fig. 3

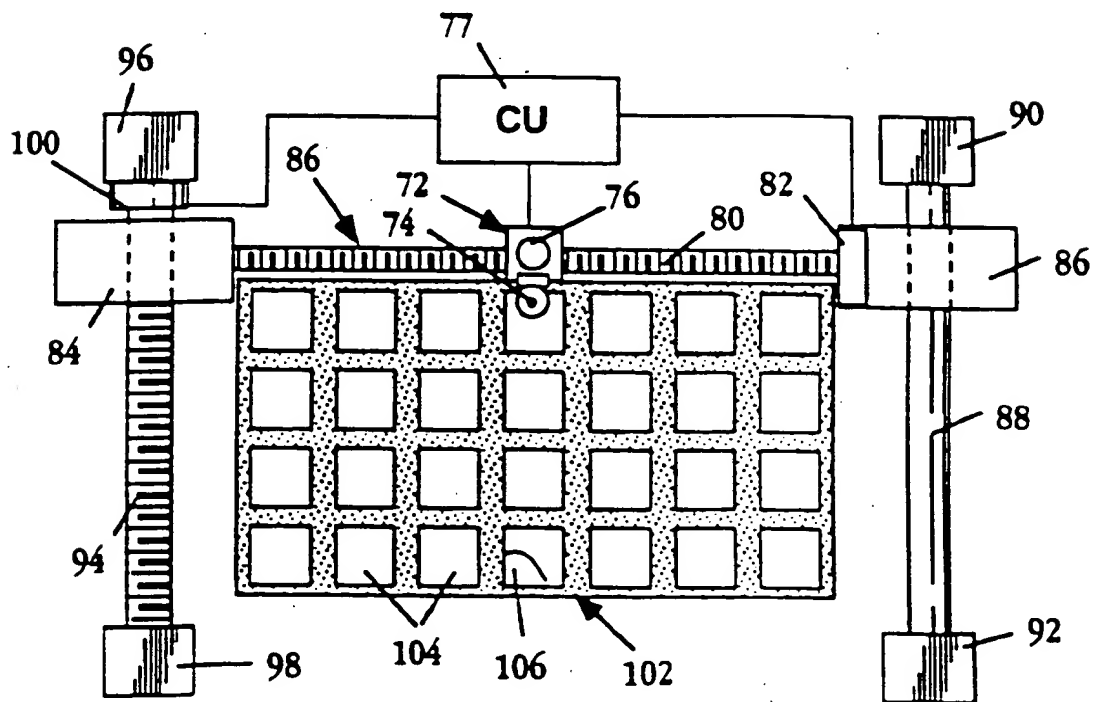


Fig. 4

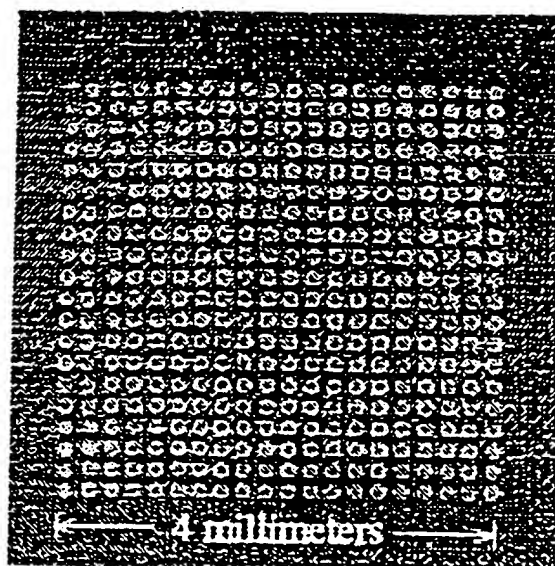


Fig. 5

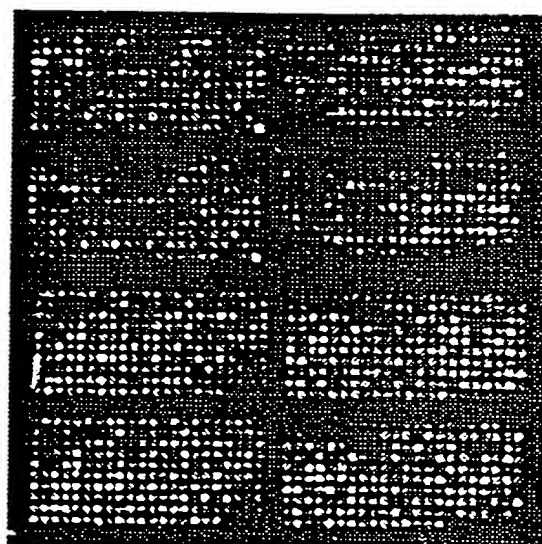


Fig. 6



Fig. 7

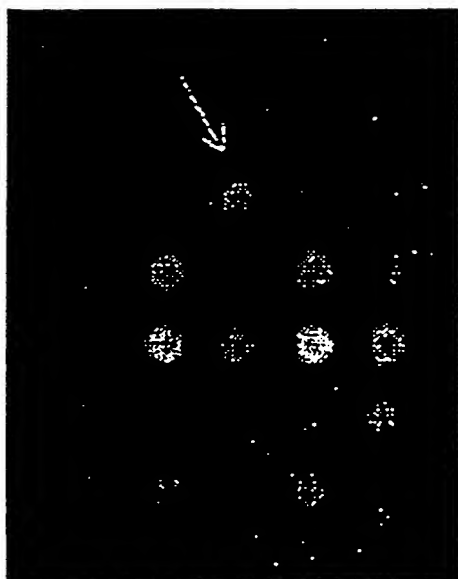


Fig. 8

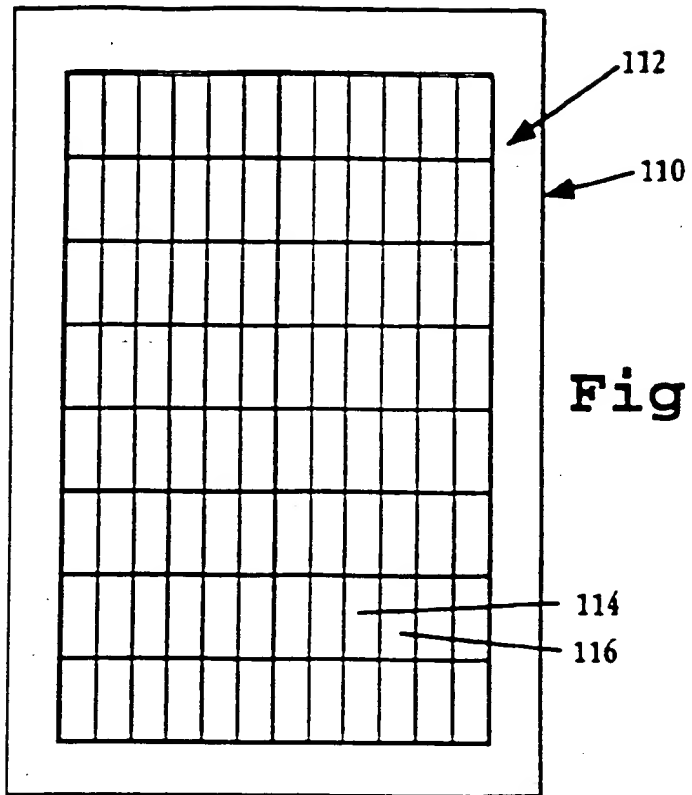


Fig. 9

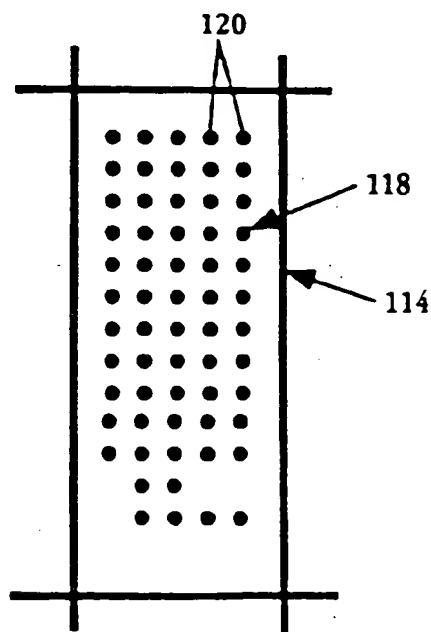


Fig. 10

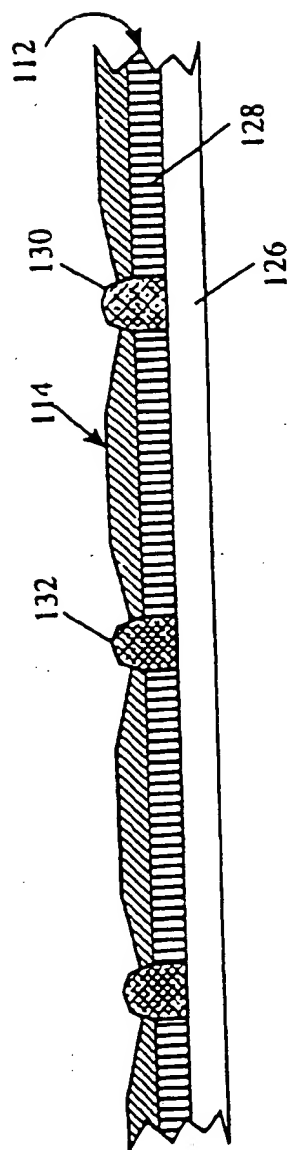


Fig. 11

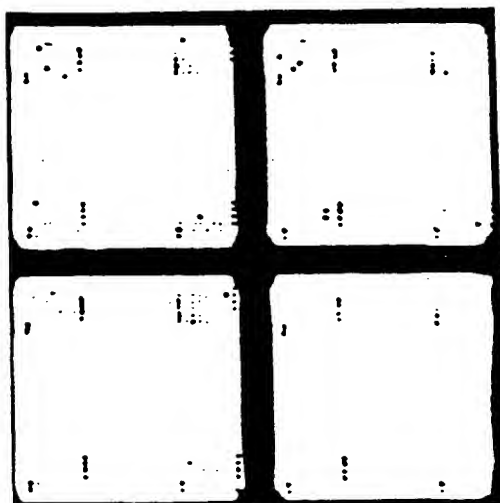


Fig. 12

METHODS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

CROSS-REFERENCE TO RELATED APPLICATION

This application is a continuation-in-part of U.S. patent application Ser. No. 08/261,388, filed Jun. 17, 1994, and now abandoned.

The United States government may have certain rights in the present invention pursuant to Grant No. HG00450 awarded by the National Institutes of Health.

FIELD OF THE INVENTION

This invention relates to a method and apparatus for fabricating microarrays of biological samples for large scale screening assays, such as arrays of DNA samples to be used in DNA hybridization assays for genetic research and diagnostic applications.

REFERENCES

- Abouzied, et al., *Journal of AOAC International* 77(2):495-500 (1994).
 Bohlander, et al., *Genomics* 13:1322-1324 (1992).
 Drmanac, et al., *Science* 260:1649-1652 (1993).
 Fodor, et al., *Science* 251:767-773 (1991).
 Khrapko, et al., *DNA Sequence* 1:375-388 (1991).
 Kuriyama, et al., *AN ISFET BIOSENSOR, APPLIED BIOSENSORS* (Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).
 Lebrach, et al., *HYBRIDIZATION FINGERPRINTING IN GENOME MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1* (Davies and Tügham, Eds.), Cold Spring Harbor Press, pp. 39-81 (1990).
 Maniatis, et al., *MOLECULAR CLONING, A LABORATORY MANUAL*, Cold Spring Harbor Press (1989).
 Nelson, et al., *Nature Genetics* 4:11-18 (1993).
 Pittung, et al., U.S. Pat. No. 5,143,854 (1992).
 Riles, et al., *Genetics* 134:81-150 (1993).
 Schena, M. et al., *Proc. Nat. Acad. Sci. USA* 89:3894-3898 (1992).
 Southern, et al., *Genomics* 13:1008-1017 (1992).

BACKGROUND OF THE INVENTION

A variety of methods are currently available for making arrays of biological macromolecules, such as arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation. This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a

porous membrane. One array includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22x22 cm area (Lebrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable. In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pittung, et al. (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatis, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the immobilized antibody and an antigen is detected using a standard ELISA calorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen (Research Products International). However, Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the hydrophobic barrier during extended high temperature incubations or in the presence of detergents, which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego Calif.) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate. These plates are

capable of containing different reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many colorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barrier elements are shallow and do not interfere with the detection step, thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

SUMMARY OF THE INVENTION

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel, and deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is applied to a selected position on each of a plurality of solid supports at each repeat cycle.

The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes a positioning structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and a dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit operates to (i)

place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) removing the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm². Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film. The grid is composed of intersecting water-impermeable grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impermeable cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impermeable grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescently-labeled cDNAs from mRNAs isolated from the two cells types, where the cDNAs from the first and second cell types are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNAs from the two cell types is added to an array of polynucleotides representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNAs to complementary-sequence polynucleotides in the array. The array is then examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNAs derived

from one of the first or second cell types give a distinct first or second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNAs derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read in conjunction with the accompanying figures.

The file of this patent contains at least one drawing executed in color. Copies of this patent with color drawing(s) will be provided by the Patent and Trademark Office upon request and payment of the necessary fee.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head constructed for use in one embodiment of the invention;

FIGS. 2A-2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from FIG. 1, in accordance with one embodiment of the method of the invention;

FIG. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

FIG. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance with the invention.

FIG. 5 shows a fluorescent image of an actual 20x20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-L-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

FIG. 6 is a fluorescent image of a 1.8 cmx1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

FIG. 7 shows the translation of the hybridization image of FIG. 6 into a karyotype of the yeast genome, where the elements of FIG. 6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

FIG. 8 shows a fluorescent image of a 0.5 cmx0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type Arabidopsis plant labeled with a green fluorophore and total cDNA from a transgenic Arabidopsis plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic Arabidopsis plant;

FIG. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

FIG. 10 shows an enlarged plan view of one of the cells in the substrate in FIG. 9, showing an array of polynucleotide regions in the cell;

FIG. 11 is an enlarged sectional view of the substrate in FIG. 9, taken along a section line in that figure; and

FIG. 12 is a scanned image of a 3 cmx3 cm microcellulose solid support containing four identical arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

DETAILED DESCRIPTION OF THE INVENTION

I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

"Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair, an effector molecule in an effector/receptor binding pair, or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Anti-ligand" refers to the opposite member of a ligand/anti-ligand binding pair. The anti-ligand may be the other of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair, the receptor molecule in an effector/receptor binding pair, or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

"Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

"Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a linear or two-dimensional array of preferably discrete regions, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about 100/cm², and preferably at least about 1000/cm². The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μ m, and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

FIG. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below. The capillary channel is formed by a pair of spaced-apart, coaxial, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in FIG. 2A. The advantages of the open channel construction of the dispenser are discussed below.

With continued reference to FIG. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in the dispenser on a support, as will be described below with reference to FIGS. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring bias, to a normal, raised position, as shown. The dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

FIGS. 2A-2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, i.e., wettable surface, such as a surface having native, bound or covalently attached charged groups. One such surface described below is a glass surface having an adsorbed layer of a polycationic polymer, such as poly-L-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, i.e., one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the dispenser tip in its raised position, as seen in FIG. 2A, where the tip is typically at least several 1-5 mm above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move rapidly toward

and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in FIG. 2B.

FIG. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size, i.e., volume, the tendency of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, e.g., the positions illustrated in FIGS. 2B and 2C.

The desired deposition volume, i.e., bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, i.e., the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in FIG. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μ m.

TABLE 1

d	Volume (nl)
20 μ m	2×10^{-3}
50 μ m	3.1×10^{-2}
100 μ m	2.5×10^{-1}
200 μ m	2

At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface,

and/or increasing the viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that the tweeter-like, open-capillary dispenser tip provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in FIG. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20–200 μm . The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20–400 μm . Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/cm². After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in FIG. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to FIG. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in FIGS. 1 and 2A–2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above.

Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5–25 μm . In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more, for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along a y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5–250 μm , and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of reagent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings

that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuation of the dispenser is then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 μ l and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried out as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

A. Multi-Cell Substrate

FIG. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8x12 rectangular array 112 of cells, such as cells 114, 116, formed on the substrate surface. With reference to FIG. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in FIG. 3.

The 96-cell array shown in FIG. 9 typically has array dimensions between about 12 and 244 mm in width and 8 and 400 mm in length, with the cells in the array having width and length dimension of $\frac{1}{2}$ and $\frac{1}{4}$ the array width and length dimensions, respectively, i.e., between about 1 and 20 in width and 1 and 50 mm in length.

The construction of substrate is shown cross-sectionally in FIG. 11, which is an enlarged sectional view taken along view line 124 in FIG. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed on the surface of the backing is a water-permeable film 128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μ m. The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to FIG. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μ m above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-imperious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In FIG. 11, defined volume samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the microarrays in the well are identical arrays of distinct biopolymers, e.g., different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

In a preferred embodiment, each microarray contains about 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . Also in a preferred embodiment, the biopolymers in each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiment, the biopolymers are polynucleotides having lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by schemes involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard calorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

FIG. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-L-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-L-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, Mo.).

To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is illustrated in Examples 1 and 2.

To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2 . In the embodiment shown in FIG. 5, the microarray contains 400 regions in an area of about 16 mm^2 , or 2.5×10^3 regions/ cm^2 . Also in a preferred embodiment, the polynucleotides in each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiment, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various *in situ* synthesis schemes.

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described by Nelson, et al. (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous to the immobilized clones on the array. For example, Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene (or genes) can be probed with a labeled mixture of a patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for

unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNAs and RNAs can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. An example of this approach is illustrated in Example 2, described with respect to FIG. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm x 2.2 cm microarrays fabricated on a single 12 cm x 18 cm sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross-contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatis, et al., 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug coagener preparations or chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

The following examples illustrate, but in no way are intended to limit, the present invention.

EXAMPLE 1

Genomic-Complexity Hybridization to DNA Microarrays Representing the Yeast *Saccharomyces cerevisiae* Genome with Two-Color Fluorescent Detection

The array elements were randomly amplified PCR (Bohlander, et al., 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA as templates (Riles, et al., 1993). The PCR was performed directly on the lambda phage lysates, resulting in an amplification of both the 35 kb lambda vector and the 5-15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250-1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ.) and concentrated by evaporation to dryness at room temperature overnight. Each of the 864 amplified lambda clones was rehydrated in 15 µl of 3xSSC in preparation for spotting onto the glass.

The microarrays were fabricated on microscope slides which were coated with a layer of poly-L-lysine (Sigma). The automated apparatus described in Section III loaded 1 µl of the concentrated lambda clone PCR product in 3xSSC directly from 96 well storage plates into the open capillary printing element and deposited ~5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove unabsorbed DNA and then treated with succinic anhydride in reduce non-specific adsorption of the labeled hybridization probe to the poly-L-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90° for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, Calif.). The six largest chromosomes were isolated in one gel slice and the ten smallest chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, Calif.). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, Ill.) with a biotinylated nucleotide analog (Dupont NEN, Boston, Mass.) for the pool containing the six largest chromosomes, and with a fluorescein conjugated nucleotide analog (BMB) for the pool containing ten smallest chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, Mass.).

Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 µl of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 µl of concentrated hybridization solution (5xSSC and 0.1% SDS) was added and all 10 µl transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1xSSC and 0.1% SDS for 5 minutes, cover slipped and scanned.

A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8x1.8

cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype of the yeast genome.

FIG. 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the six largest yeast chromosomes. A green signal indicates that the lambda clone insert comes from one of the ten smallest yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, et al.), allowing for the automatic generation of the color karyotype shown in FIG. 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the six largest chromosomes are mainly red while the ten smallest chromosomes are mainly green, thus matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into positions on the physical map of the yeast genome.

EXAMPLE 2

Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

Twenty-four clones containing cDNA inserts from the plant *Arabidopsis* were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3xSSC. The cDNA clones were spotted on poly-L-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone representing a transcription factor HAT4, which had previously been used to create a transgenic line of the plant *Arabidopsis*, in which this gene is present at ten times the level found in wild-type *Arabidopsis* (Schemm, et al., 1992).

Total poly-A mRNA from wild type *Arabidopsis* was isolated using standard methods (Maniatis, et al., 1989) and reverse transcribed into total cDNA, using a fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of *Arabidopsis* where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lysamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization

reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. FIG. 8 shows the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic *Arabidopsis* appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of *Arabidopsis* but not detectably expressed in wild type *Arabidopsis*, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic *Arabidopsis* and the relative lack of expression of the transcription factor in the green-labeled wild type *Arabidopsis*.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

EXAMPLE 3

Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barner elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10xSSC and allowed to dry. As shown in FIG. 12, 192 M13 clones, each with a different yeast insert were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization, while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of each array are positive controls for the colorimetric detection step.

The oligonucleotides were labeled with fluorescein, which was detected using an anti-fluorescein antibody conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II) attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable dummies on the nitrocellulose which can be used to automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be clear that various changes and modification may be made without departing from the invention.

We claim:

1. A method of forming a microarray of discrete analyte-assay regions on a solid support, where each discrete region in the microarray has a selected, analyte-specific reagent, said method comprising,

(a) loading an aqueous solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel adapted to hold a quantity of the reagent solution and having a tip region at which the solution in the channel forms a meniscus,

(b) tapping the tip of the dispensing device against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume between 0.002 and 2 nl of solution on the surface, and

(c) repeating steps (a) and (b) until said microarray is formed.

2. The method of claim 1, wherein the reagents used to form the discrete regions in the microarray are distinct

nucleic acid strands and wherein steps (a) and (b) are repeated until the microarray has about 100 or more discrete regions of distinct nucleic acid strands per cm^2 of solid support.

3. The method of claim 1, wherein the reagents used to form the discrete regions in the microarray are distinct nucleic acid strands and wherein steps (a) and (b) are repeated until the microarray has about 1000 or more discrete regions of distinct nucleic acid strands per cm^2 of solid support.

4. The method of claim 2, wherein the channel is open-sided.

5. The method of claim 3, wherein the channel is open-sided.

6. The method of claim 4, wherein the volume is between 0.002 and 0.25 nl.

7. The method of claim 5, wherein the volume is between 0.002 and 0.25 nl.

• • • • •

nature *genetics*

volume 14 no. 4

december 1996

**DNA chips,
diagnostics
and genomics**

**Rieger
syndrome**

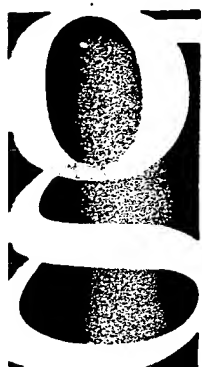
**QTLs and
epistasis**

**A BRCA1-
binding
protein**



LIBRARY
NBERG BLDG 11TH FL
GUSTAVE L. LEVY PL
YORK NY 10029-6504

00/003/0009



Nature Genetics

Editor
Levin Davies

Assistant Editors
Aune Goodman
Lette Phimister

Production Editor
Stuart Griffith

Assistant Production Editor
Ken Krattenmaker

Editorial Assistant
Anelle Bolden

Washington Bureau Chief
Barbara J. Culliton

Editorial Office
35 National Press Building
Washington DC 20045
Tel: (202) 626-2513
Fax: (202) 626-0970
Email: natgen@naturedc.com

WWW: genetics.nature.com



Cover art: Ken Krattenmaker

editorial To affinity . . . and beyond! edit 367

news & views Who's afraid of epistasis? news 371

Wayne N Frankel & Nicholas J Schork

Meiotic nondisjunction does the two-step news 374

Terry Orr-Weaver

Flood warning — resistance genes unleashed news 376

Richard Michelmore

correspondence Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence 380

A H Paterson, T-H Lan, K P Reischmann, C Chang, Y-R Lin, S-C Liu, M D Burrow, S P Kowalski, C S Katsar, T A DelMonte, K A Feldmann, K F Schertz & J F Wendel

Non-canonical introns are at least 10⁹ years old 383

H-J Wu, P Gaubier-Comella, M Delseny, F Grellet, M Van Montagu & P Rouzé

Val92Met variant of the melanocyte stimulating hormone receptor gene 384

X Xu, M Thörnwall, L-G Lundin & V Chhajlani

progress Genes responsible for human hereditary deafness: symphony of a thousand 385

Christine Petit

articles Cloning and characterization of a novel bicoid-related homeobox transcription factor gene, RIEG, involved in Rieger syndrome 392

E V Semina, R Reiter, N J Leysens, W L M Alward, K W Small, N A Datson, J Siegel-Bartelt, D Bierke-Nelson, P Bitoun, B U Zabel, J C Carey & J C Murray

Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II 400

N E Lamb, S B Freeman, A Savage-Austin, D Pettay, Lisa Taft, J Hersey, Y Gu, J Shen, D Saker, K M May, D Avramopoulos, M B Petersen, A Hallberg, M Mikkelsen, T J Hassold & S L Sherman news

Spontaneous X chromosome MI and MII nondisjunction events in *Drosophila melanogaster* oocytes have different recombinational histories 406

K E Koehler, C L Boulton, H E Collins, R L French, K C Herman, S M Laceyfield, L D Madden, C D Schuetz & R S Hawley news

Suppression of the novel growth inhibitor p33^{ING} promotes neoplastic transformation 415

I Garkavtsev, A Kazarov, A Gudkov & K Riabowol

Nature Publishing Co.
375 Park Avenue South
11th floor
New York, NY 10010-1707
Tel: (212) 726-9200
Fax: (212) 696-9606

President-Publisher
Mary Wathnam

Vice President Sales
Manon Delaney

Vice President Marketing
James A. Skowrenski

American Advertising Sales
Manager
Sande T. Giaccone (New York)

European Advertising Sales
Manager
Kathryn Wayman (London)

Classified Advertising Sales
Manager
Erika A. Simon (New York)
Mike Grant (London)

Assistant Classified Sales
Manager
Benjamin Crowe (New York)

Production & Information
Systems Director
Nick Kemp

Circulation Manager
Moira Musto (New York)
Nic Harman (London)

Group Marketing Manager
Anna Dzurenda



Macmillan Magazines Ltd
Orders South
Rivian Street
London N1 9XW
Tel: 44 (0)171 833 4000
Fax: 44 (0)171 843 4596

Managing Director
Ray Barker

Publishing Director
Andy Sutherland

Editor-in-Chief,
Nature publications
Philip Campbell

Art Director
Jane Walker

Nature Japan KK
Nishi-Mitsuke Bldg
6 Ichigaya Tamachi
Nishjuku-ku
Tokyo 162
Telephone 03 3267 8751
Fax 03 3267 8746

Publisher
David Swinbanks

articles

A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants 421
D Leister, A Ballvora, F Salamini & C Gebhardt

Identification of a RING protein that can interact *in vivo* with the BRCA1 gene product 430
L C Wu, Z W Wang, J T Tsan, M A Spillman, A Phung, X L Xu, M-C W Yang, L-Y Hwang, A M Bowcock & R Baer

Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis 441
J G Hacia, L C Brody, M S Chee, S P A Fodor & F S Collins

Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy 450
D D Shoemaker, D A Lashkari, D Morris, M Mittmann & R W Davis

letters

Use of a cDNA microarray to analyse gene expression patterns in human cancer 457
J DeRisi, L Penland & P O Brown (Group 1); M L Bittner, P S Meltzer, M Ray, Y Chen, Y A Su & J M Trent (Group 2)

Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis 461
I Perrault, J M Rozet, P Calvas, S Gerber, A Camuzat, H Dollfus, S Châtelain, E Souied, I Ghazi, C Leowski, M Bonnemaïson, D Le Paslier, J Frézal, J-L Dufer, S Pittler, A Munnich & J Kaplan

Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse 465
R J A Fijneman, S S de Vries, R C Jansen & P Demant

Gene interaction and single gene effects in colon tumour susceptibility in mice 468
T van Wezel, A P M Stassen, C J A Moen, A A M Hart, M A van der Valk & P Demant

A major quantitative trait locus influences hyperactivity in the WKHA rat 471
M-P Moisan, H Courvoisier, M-T Bihoreau, D Gauguier, E D Hendley, M Lathrop, M R James & P Mormède

An H-YD^b epitope is encoded by a novel mouse Y chromosome gene 474
A Greenfield, D Scott, D Pennisi, I Ehrmann, P Ellis, L Cooper, E Simpson & P Koopman

Homozygosity mapping of Hallervorden-Spatz syndrome to chromosome 20p12.3-p13 479
T D Taylor, M Litt, P Kramer, M Pandolfo, L Angelini, N Nardocci, S Davis, M Pineda, H Hattori, P J Flett, M R Cilio, E Bertini & S J Hayflick

Identification of BTG2, an antiproliferative p53-dependent component of the DNA damage cellular response pathway 482
J-P Rouault, N Falette, F Guéhenneux, C Guillot, R Rimokh, Q Wang, C Berthet, C Moyret-Lalle, P Savatier, B Pain, P Shaw, R Berger, J Samarut, J-P Magaud, M Ozturk, C Samarut & A Puisieux

correction/errata See pages 487-488

classified See back pages

Use of a cDNA microarray to analyse gene expression patterns in human cancer

Joseph DeRisi¹*, Lolita Penland² & Patrick O. Brown² (Group 1); Michael L. Bittner³*, Paul S. Meltzer³, Michael Ray³, Yidong Chen³, Yan A. Su³ & Jeffrey M. Trent³ (Group 2)

The development and progression of cancer¹⁻³ and the experimental reversal of tumorigenicity^{4,5} are accompanied by complex changes in patterns of gene expression. Microarrays of cDNA provide a powerful tool for studying these complex phenomena⁶⁻⁸. The tumorigenic properties of a human melanoma cell line, UACC-903, can be suppressed by introduction of a normal human chromosome 6, resulting in a reduction of growth rate, restoration of contact inhibition, and suppression of both soft agar clonogenicity and tumorigenicity in nude mice^{4,5,9}. We used a high density microarray of 1,161 DNA elements to search for differences in gene expression associated with tumour suppression in this system. Fluorescent probes for hybridization were derived from two sources of cellular mRNA [UACC-903 and UACC-903(+6)] which were labelled with different fluors to provide a direct and internally controlled comparison of the mRNA levels corresponding to each arrayed gene. The fluorescence signals representing hybridization to each arrayed gene were analysed to determine the relative abundance in the two samples of mRNAs corresponding to each gene. Previously unrecognized alterations in the expression of specific genes provide leads for further investigation of the genetic basis of the tumorigenic phenotype of these cells.

DNA microarrays, containing 1,161 total elements, including 870 different cDNAs and controls⁹⁻¹¹ (see Methods), were printed robotically onto a glass microscope slide in four quadrants covering an area of about 1 cm² (Fig. 1). We prepared fluorescent cDNA probes using total poly (A)⁺ mRNA from UACC-903 cells and UACC-903(+6) cells by labelling with a green and red fluor, respectively. A mixture of the two fluorescently labelled probes was hybridized to the DNA microarray. This comparative hybridization method, coupled with the doping of synthetic standards and an estimation of statistically significant deviation for local background variance allowed a direct and quantitative comparison of the relative abundance of individual DNA sequences in this complex sample⁶⁻⁸. We added a set of synthetic poly (A)⁺-tailed 'mRNAs' to the purified mRNA from each cell line as internal standards to assist in quantitation and estimation of experimental variation introduced during labelling and reading. Targets complementary to these standards were included, in duplicate, on the microarray. Based on these standards, mRNA expression levels were determined for each gene.

spend to genes preferentially expressed in the tumorigenic UACC-903 cell line, and the reddish spots correspond to genes preferentially expressed in the non-tumorigenic UACC-903(+6) cell line. Genes expressed at approximately equal levels in the two cell lines appear yellow or brown. A portion of the array at higher magnification highlights the diverse pattern of differential expression observed (Fig. 2b). In Fig. 2c, rectangles corresponding to specific array elements are coloured to reproduce the hue and intensity of the fluorescent signal at each element. The hybridization signals from a duplicated set of genes are shown juxtaposed, to illustrate the reproducibility of the hybridization signals for each gene.

To address the possibility that an apparent difference in expression might result from experimental variables unrelated to the difference in chromosomal composition between the two cell lines, we examined the variance in expression for 90 'housekeeping' genes. We selected these genes based on the assumption that they would not be differentially expressed between the two cell lines. The averaged red/green ratio for this subset of genes was 1.13. The averaged red/green ratio for the set of five internal standards was 0.97 ($n = 10$). The variability in the expression level of the housekeeping genes probably overestimates the experimental variability in measuring differential expression. As a conservative standard, an absolute fluorescent signal (red or green) with an intensity greater than that observed at the control array elements containing total human genomic DNA was considered to represent specific hybridization. Gene-specific hybridization was therefore only considered significantly different between samples if the following two criteria were met: i) the signal intensity (green or red) exceeded this threshold; and ii) the logarithm of the red/green fluorescence signal ratio differed by ≥ 3 S.D. from the mean logarithm of this ratio for the 'housekeeping' gene panel (that is, ratios <0.52 or >2.4).

By these criteria, mRNA levels for 15/870 (1.7%) genes were significantly diminished, while the mRNA levels for 63/870 (7.3%) genes were significantly increased association with suppression of tumorigenicity by introduction of chromosome 6. To test the reliability of microarray hybridization results in identifying differentially expressed genes, we analysed 16 genes by northern analysis. In each case, the results of northern analysis corroborated the differential gene expression identified by microarray hybridization (Fig. 3).

Significant differences in expression between the two cell lines identified several genes as candidate features of the tumorigenic phenotype of the melanoma cells. For example, among genes detected with significantly higher expression in the tumorigenic cells was the human tryptophan receptor (TRP1/melanoma antigen gp75). This abundant glycoprotein in melanocytic cells is a melanosome membrane protein^{12,13}. Its expression is reduced when melanocytes are induced to differentiate by treatment with retinoic acid. Also expressed at a significantly higher level was a variant of the mRNA encoding myosin.

¹Howard Hughes Medical Institute.
²Department of Biochemistry, Stanford University Medical Center, Stanford, California 94305, USA
³Laboratory of Cancer Genetics, National Center for Human Genome Research, National Institutes of Health, Bethesda, Maryland 20892, USA

*J.D. & M.L.B. contributed equally to this work.

Correspondence should be addressed to P.O. or J.T.

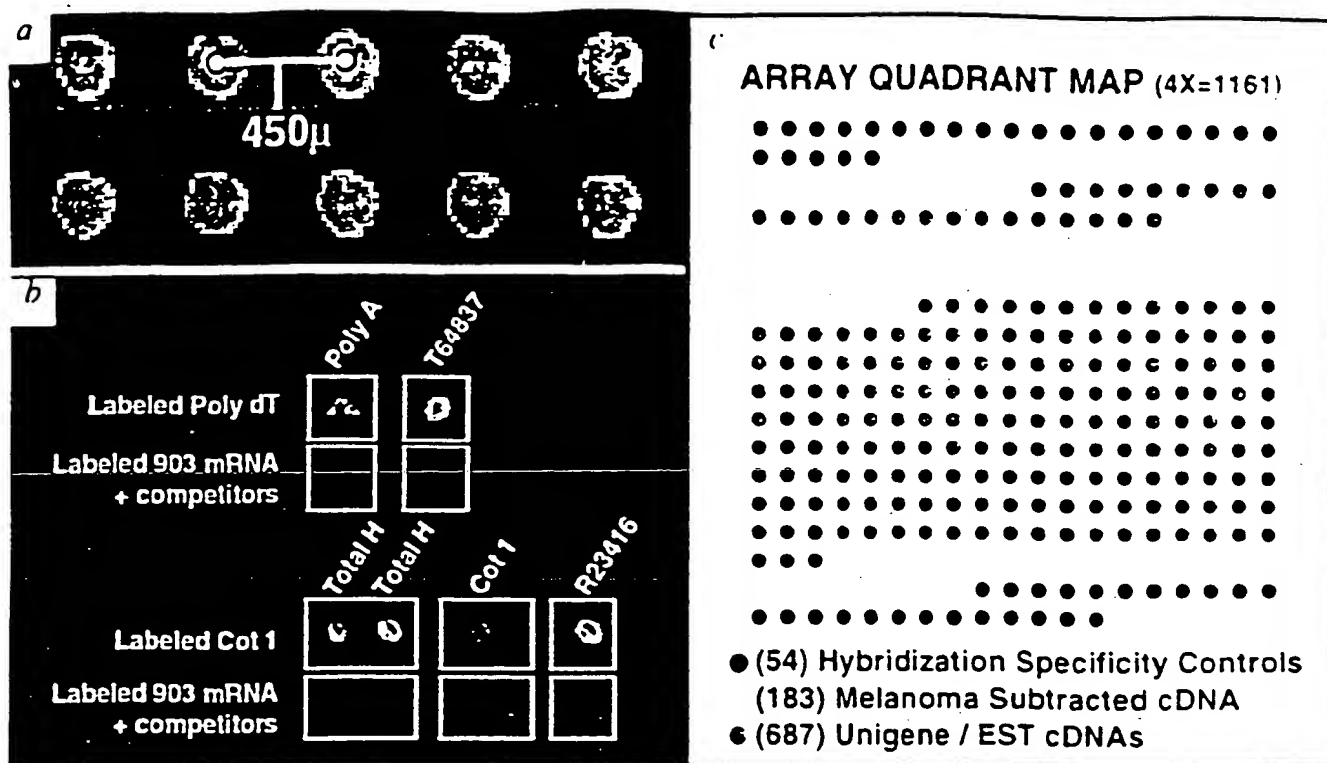


Fig. 1 Properties of cDNA microarrays. **a.** A fluorescent scan of DNA printed onto a poly-lysine coated slide. The DNA is stained with a DNA-specific fluorescent dye, YOYO. The center-to-center spacing of adjacent spots is 450 μ m, allowing the potential for up to 10,000 spots/2.54 X 7.62 cm microscope slide. **b.** Efficient blocking of hybridization to DNA repeats. Hybridization of fluorescein-labelled poly (dT)* to arrays in the absence of competitor produces strong hybridization to immobilized poly (dA)* as well as to some cDNAs, such as the EST T64827 shown. Rhodamine-labelled cDNA (red) from the UACC-903 cell line hybridized in the presence of poly (dA)* blocker shows little if any signal at either site (Total H = total human). Similarly, hybridization with fluorescein-labelled Cot1 DNA in the absence of competitor produces bright signal on immobilized Cot1 DNA, total human DNA and at some cDNA elements (presumed to contain highly repeated sequences, such as R23416); while Rhodamine-labelled cDNA (red) from the UACC-903 cell line produces little if any signal at these locations when hybridized in the presence of excess unlabelled poly (dA)*, and human Cot1 DNA. The absence of signal at some cDNA locations following UACC-903 cDNA hybridizations also indicates that the PCR-amplified, plasmid vector sequences at all cDNA targets do not contribute significant hybridization signal. **c.** Schematic of the array organisation. Robotic printing from 96 well microtiter trays was carried out with 4 print heads, spaced to fit into 4 adjacent microliter wells. This maps the contents of each tray into four separate quadrants on the glass slide. A colour-coded map of the general distribution of target types in each of the resulting quadrants is shown.

els were elevated by the addition of a normal chromosome 6 (17 genes) are known to be activated by IFN- γ , a cardinal proinflammatory cytokine that, among other activities, induces expression of the gene products of the MHC class II locus. For example, the mRNA encoding monocyte chemoattractant protein 1 (MCAF/MCP1), a cytokine that induces monocyte chemotaxis and activation^{15,16}, was more than 10-fold less abundant in the tumorigenic cell line. In the skin, MCP1 is critical in the regulation of cutaneous monocyte trafficking¹⁶⁻¹⁷, and elevated expression plays a role in suppression of tumour growth and metastasis¹⁹⁻²¹. The mechanism by which these interferon- γ regulated genes are induced in UACC-903 cells by transfer of a normal chromosome 6 remains to be determined. It is worth noting, however, that the interferon- γ receptor gene is localized to the distal long arm of human chromosome 6.

Finally, several genes that showed >10-fold higher expression in the suppressed UACC-903(+6) cells have previously been recognized in other models of tumour suppression. Most notably, there was elevated expression of the mRNA encoding WAF1 (p21), a key mediator of tumour suppression by p53 (ref. 18). The p21 protein had previously been identified as a melanoma differentiation-associated antigen (termed mda-6)^{19,20}. In melanoma cell lines suppressed for metastasis by the introduction of chromosome 6, expression of WAF1 (p21) mRNA and protein correlates inversely with

These results provide a wide view of the diverse systems that are altered in this model system of tumorigenicity, and focus attention on specific gene products and pathways that may be of particular importance in this tumour type.

Our ability to classify human cancers in a way that reflects the underlying molecular pathology or that anticipates their potential for progression or response to treatment, remains primitive. Using cDNA microarrays to define alterations in gene expression associated with a specific cancer may be an efficient way to uncover clues to the specific molecular derangements that contribute to its pathogenesis and thus identify potential targets for therapeutic intervention. Moreover, recognition of pathognomonic alterations in gene expression might provide a basis for improved diagnosis and molecular classification of cancers and thus allow selection of the most appropriate therapeutic strategies.

Public databases of human expressed gene sequences contain partial sequences of at least 40,000 different human genes¹¹, and efforts to develop a human transcript map have developed rapidly²¹. Based on the high yield of information obtained using an array of <1,000 different genes, a more comprehensive survey of gene expression patterns, using a more complete array of human genes, will likely provide a rich source of new and useful insights into human biology and a deeper understanding of the gene pathways involved in the

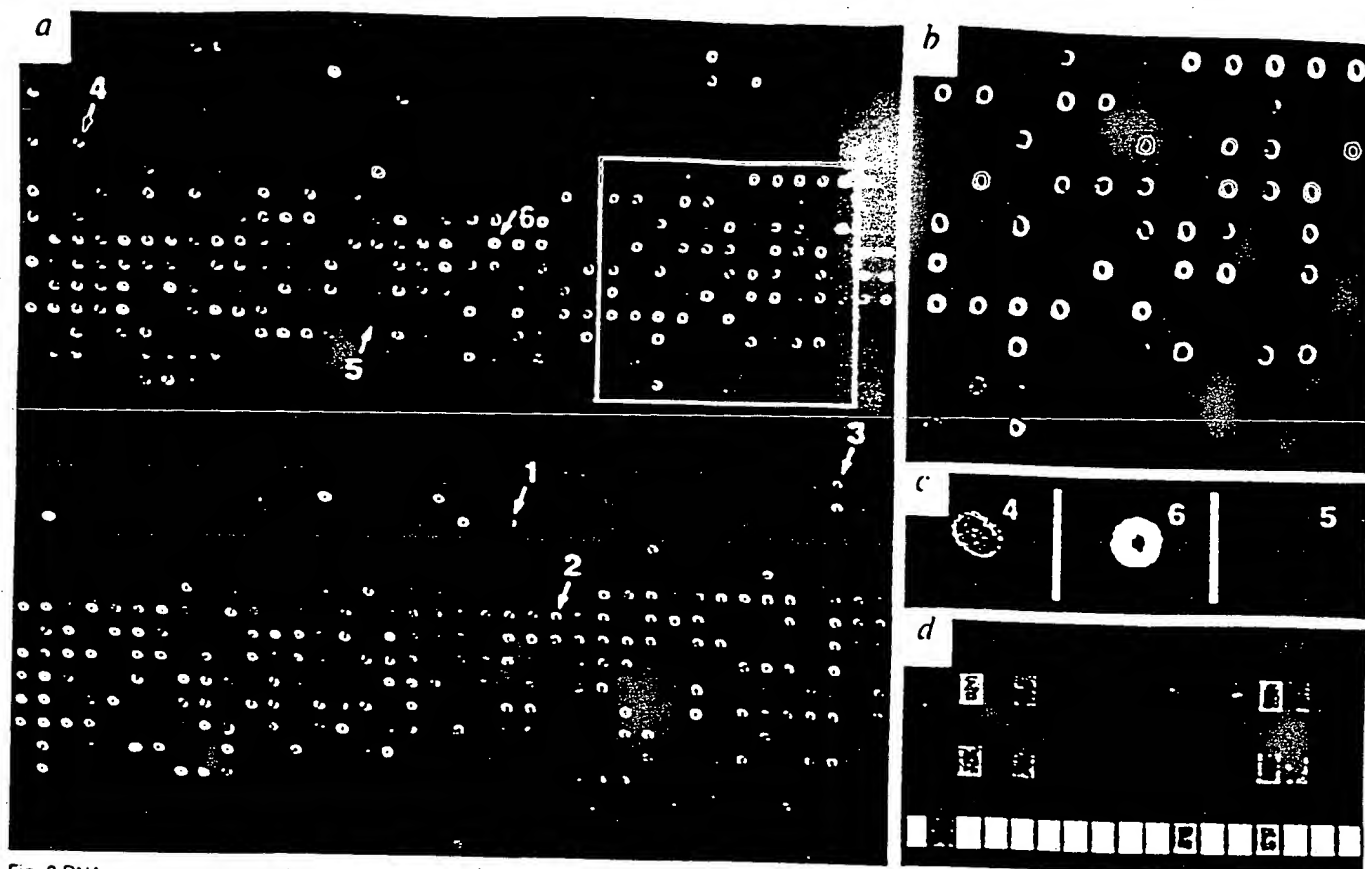


Fig. 2 DNA microarray analysis of changes in gene expression between the tumorigenic cell line, UACC-903, and its non-tumorigenic derivative, UACC903(+6), derived by introduction of a normal chromosome 6. **a**, A ratio image of the results of simultaneous hybridization of Rhodamine 110-labelled cDNA (green) from UACC-903 and Cy3-labelled cDNA (orange-red) from UACC-903(+6) to a microarray. To produce this image, the scan images corresponding to each fluorescent probe were combined as the appropriate colour channels in a single image. Arrows indicate the location within the array of the corresponding genes analysed by northern blotting (Fig. 3). **b**, A magnified image of the area of the array boxed in white in (a). **c**, Magnified image of three cDNAs identified by arrows in (a), representing the cDNAs for: left, *MCAF/MCP-1* (*r/g* ratio > 10); centre, β -actin (*r/g* ratio 1.04); and right, *u-1-antichymotrypsin* (*r/g* ratio 0.2) [see Fig. 3]. **d**, simplified representation of ratio hybridization results. Quantitative fluorescence intensity data is extracted from each array target. The average target colour ratio determines the hue of each box and the average intensity determines the brightness of each box. In this image, the order of the boxes corresponds to their original order in the microliter plate from which they were printed. Duplicate printings of the same plate can be examined side by side, as in the first two rows shown here, to assess reproducibility of the hybridization results (see text). Numbered arrows indicate the location within the array corresponding to genes analysed by northern blotting in Fig. 3.

Methods

Generation of microarrays, hybridization, scanning. The preparation of coated microscope slides and subsequent robotic printing of DNA was carried out in a manner similar to that described [1]. Briefly, pre-cleaned glass slides were treated with poly-L-lysine solution (Sigma) to form an adhesive surface for printing. PCR products, purified by ethanol purification, were resuspended in 3x SSC. A custom built arraying robot picked up and deposited small volumes (~5 nanoliters) of DNA onto the slides. After printing, the slides were washed in a 0.2% SDS solution. The remaining bound DNA was denatured by submerging the slides in 95 °C distilled water for 2 min followed by a brief wash with 95% ethanol. DNA was UV crosslinked to the slides (Stratagene Stratalinker, 60 mJ). To prevent non-specific probe binding, the slides were blocked by rinsing in a solution of 70 mM succinic anhydride dissolved in 0.1 M boric acid pH 8.0, containing 35% 1-methyl-2-pyrrolidinone (Aldrich). Additional protocols and parts list pertaining to microarray fabrication can be obtained from rtm@cmgm.stanford.edu or rtm@cmgm.stanford.edu.

Purified, labelled cDNA was resuspended in 11 μ l of 3.5x SSC containing 4 μ g of poly (dA)⁺ DNA, 2.5 μ g *E. coli* tRNA, 4 μ g of human Cot1 DNA (Gibco BRL), and 0.3 μ l of 10% SDS. Prior to hybridization, the solution was boiled for 2 min then allowed to cool to room temperature. Hybridization was carried out at

by N. Ziv. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. Data was collected at a maximum resolution of 9 microns/pixel with 12 bits of depth.

Probe preparation and labelling. RNA was extracted from cells using the Triazol reagent (LTI Inc.), following the manufacturer's directions. cDNA probes were synthesized from singly oligo dT-selected (Pharmacia) mRNA pools. Fluorescently labelled cDNA was prepared from mRNA by oligo dT-primed polymerization using SuperScript II reverse transcriptase (LTI Inc.). The pool of nucleotides in the labelling reaction was 0.5 mM dGTP, dATP and dCTP and 0.2 mM dTTP. Fluorescent nucleotides, Rhodamine 110 dUTP (Perkin Elmer Cetus) or Cy3 dUTP (Amersham), were present at 0.1 mM. Probes were purified by gel chromatography (BioSpin 6/BioRad) and ethanol precipitation.

Selection of cDNA elements and generation of control templates. Synthetic cDNAs were prepared by cloning random *Bam*HI and *Hind*III ended fragments of *E. coli* DNA in the vector pSP64 poly (A)⁺ (Promega), linearizing isolated plasmid DNA with *Eco*RI and synthesizing poly (A)⁺ tailed RNA complementary to the insert from the resident SP6 promoter (Promega). Prior to use, the synthesized RNAs were columned and

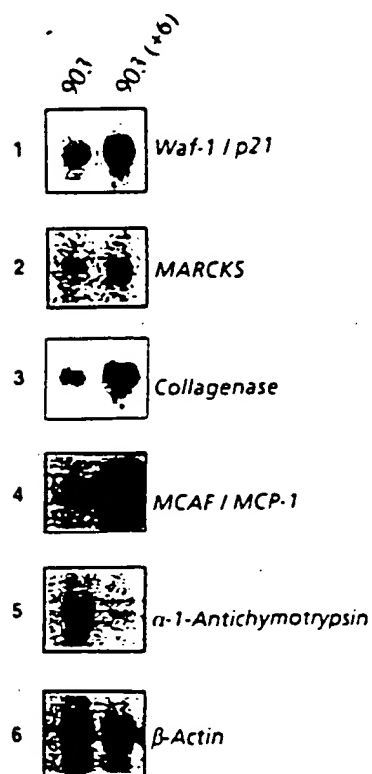


Fig. 3 Northern hybridization substantiating the consistency of the cDNA microarray results. Corresponding locations within the cDNA microarray illustrated in Fig. 2a are provided for 1) *Waf-1/p21*; 2) *MARCKS*; 3) *collagenase*; 4) *MCAF/MCP-1*; 5) α -1-*antichymotrypsin*; and 6) β -*actin*. The signal detected by a radio-labelled β -*actin* probe represents a control for loading variance, with a red/green ratio observed on the cDNA microarray (Fig. 2a,c) for β -*actin* of 1.04.

to the UniGene EST clustering system^{21,22}. The second largest group of clones consisted of 183 sequenced cDNA clones generated by subtraction of cDNA from the chromosome-6 suppressed non-tumorigenic UACC-903 (+6) cell line with cDNA from its parental tumorigenic cell line UACC-903 (ref. 9). Approximately 100 additional genes (total 870 genes arrayed) were obtained from EST libraries on the basis of their expression pattern (tissue specific, and so on). Each array included the following hybridization controls: plasmid vector, lambda, ϕ X174 phage, total human DNA, human *Col1* DNA, and poly (A)⁺. The synthetic standards used for normalization of signals in each wavelength were also arrayed. Controls were included in

each quadrant of the array to assess the reproducibility of the hybridization signal. Two plates of cDNA clones (derived from the UACC-903 subtracted library) were also arrayed in duplicate. Fidelity of the Unigene array relative to dbEST was tested by sequencing of a random sample of 11 clones used for microarray construction. All sequences were identical with the

corresponding dbEST entries. Additionally, each microarrayed cDNA from the UACC-903 subtracted library was sequenced. A listing of cDNAs comprising this microarray which were derived from the Unigene and 'housekeeping' panel can be obtained from <http://www.nih.gov/DIR/LCG/ARRAY/expn.html>.

Northern blot analysis. Total RNA, 10 μ g per lane, was electrophoresed in 1.2% agarose-formaldehyde gels and transferred onto nylon membrane (Hybond-N⁺, Amersham) by capillary blotting overnight. For DNA probes insert fragments from the Soares INIB cDNA library¹⁰ were obtained by vector PCR for *p21*, *MARCKS*, α -1-*antichymotrypsin* and β -*actin*. Probes for fibroblast collagenase and MCAF/MCP-1 were isolated from a UACC-903(+6) enriched cDNA library⁹ with all probes labelled by random priming. Filters were washed to a stringency of 0.1x SSC at 42 °C for 20 min.

Web sites. <http://cmgm.stanford.edu/phrown> for protocols and parts list pertaining to microarray fabrication, <http://www.ncbi.nlm.nih.gov/DIR/LCG/ARRAY/expn.html> for a listing of cDNAs comprising this microarray which were derived from the Unigene and 'housekeeping' panel.

Acknowledgements

Work in P.O.B.'s laboratory is supported in part by the Howard Hughes Medical Institute and National Center for Human Genome Research (HG00450). We would like to acknowledge the excellent technical and graphic assistance of X. He, T. Hofmann, Y. Jiang, J. Leenders, D. Lam and B. Walker. J.D. was supported by NIH grant 2T32BM07276-21. P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

Received 15 October; accepted 8 November, 1996.

- Vogelstein, B. & Kinzler, K.W. The multistep nature of cancer. *Trends Genet.* 9, 138-141 (1993).
- Weinberg, R.A. The molecular basis of oncogenes and tumor suppressor genes. *Ann. NY Acad. Sci.* 758, 331-338 (1995).
- Levine, A.J. The tumor suppressor genes. *Annu. Rev. Biochem.* 62, 623-651 (1993).
- Trent, J.M. et al. Tumorigenicity in human melanoma cell lines controlled by introduction of human chromosome 6. *Science* 247, 568-571 (1990).
- Su, Y. et al. Reversion of monochromosome-mediated suppression of tumorigenicity in malignant melanoma by retroviral transduction. *Cancer Res.* 56, 3186-3191 (1996).
- Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470 (1995).
- Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639-645 (1996).
- Schena, M. et al. Parallel human genome analysis: microarray-based expression of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93, 10539-10544 (1996).
- Ray, M.E., Su, Y.A., Meltzer, P.S. & Trent, J.M. Isolation and characterization of genes associated with chromosome-6 mediated tumor suppression in human malignant melanoma. *Oncogene* 12, 2527-2533 (1996).
- Soares, M.B. et al. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* 91, 9228-9232 (1994).
- Boguski, M.S. & Schuler, G.D. ESTablishing a human transcript map. *Nature Genet.* 10, 369-371 (1995).
- Vijayasaradhi, S., Doskoch, P.M., Wolchok, J. & Houghton, A.N. Melanocyte differentiation marker gp75, the brown locus protein, can be regulated independently of tyrosinase and pigmentation. *J. Invest. Dermatol.* 105, 113-119 (1995).
- Vijayasaradhi, S., Xu, Y., Bouchard, B. & Houghton, A.N. Intracellular sorting and targeting of melanosomal membrane proteins: identification of signals for sorting of the human brown locus protein, gp75. *J. Invest. Dermatol.* 130, 807-820 (1995).
- Nakao, J. et al. Expression of proteolipid protein gene is directly associated with secretion of a factor influencing oligodendrocyte development. *J. Neurochem.* 64, 2396-2403 (1995).
- Graves, D.T., Barnhill, R., Galanopoulos, T. & Antonades, H.N. Expression of monocyte chemotactic protein-1 in human melanoma in vivo. *Am. J. Pathol.* 140, 9-14 (1992).
- Kristensen, M.S., Deleuran, B.W., Larsen, C.G., Thstrup-Pedersen, K. & Paludan, K. Expression of monocyte chemotactic and activating factor (MCAF) in skin related cells. A comparative study. *Cytokine* 5, 520-524 (1993).
- Huang, S., Xie, K., Singh, R.K., Gulman, M. & Bar-Eli, M. Suppression of tumor growth and metastasis of murine renal adenocarcinoma by syngeneic fibroblasts genetically engineered to secrete the JE/MCP-1 cytokine. *J. Interferon Cytokine Res.* 15, 655-665 (1995).
- El-Deiry, W.S. et al. WAF1, a potential mediator of p53 tumor suppression. *Cell* 75, 817-825 (1993).
- Miele, M.E. et al. Metastasis suppressed, but tumorigenicity and local invasiveness unaffected, in the human melanoma cell line MeLuSo after introduction of human chromosomes 1 or 6. *Mol. Carcinog.* 15, 284-299 (1996).
- Jiang, H. et al. The melanoma differentiation-associated gene mda-6, which encodes the cyclin-dependent kinase inhibitor p21, is differentially expressed during growth, differentiation and progression in human melanoma cells. *Oncogene* 10, 1855-1864 (1995).
- Schuler, G.D. et al. A gene map of the human genome. *Science* 274, 540-546 (1996).
- Lennon, G., Aufray, C., Polymeropoulos, M. & Soares, M.B. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* 33, 151-152 (1996).

W 1088-9051

GENOME RESEARCH

July 1996

Volume 6 Number 7

INCLUDING
PCR

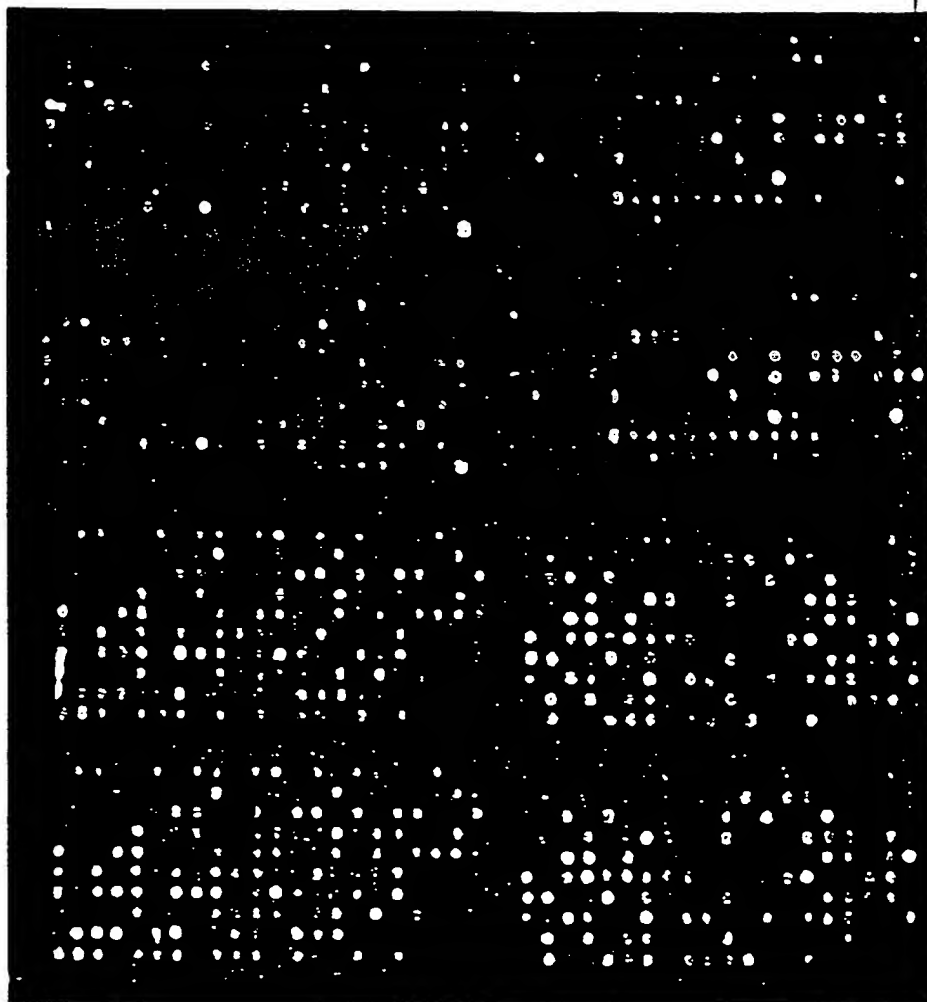
IBD Mapping in Livestock

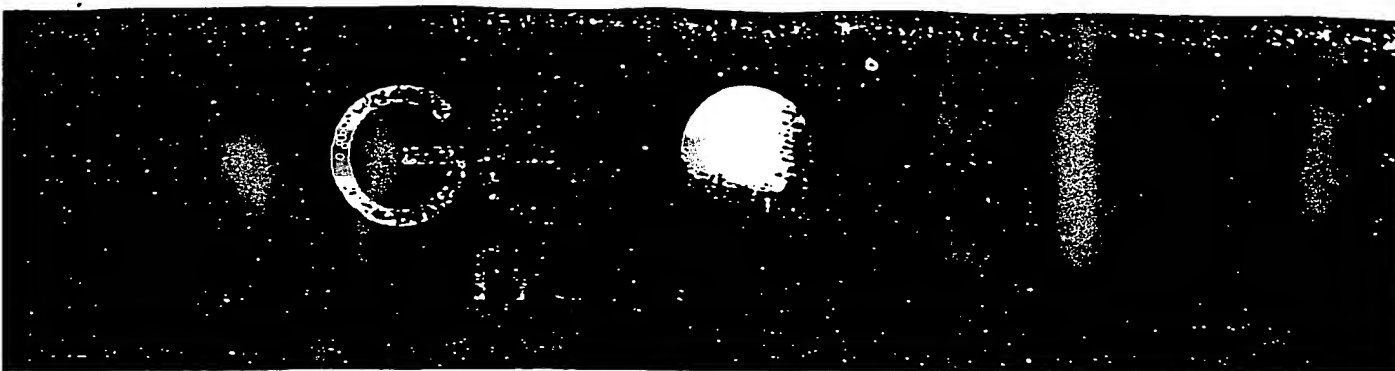
Sequence of 500-kb
Rhizobium Replicon

Human Y Chromosome
Haplotypes

BAC Mapping of
Extrachromosomal Structure

DNA Microarray System





Advertise in
Genome Research
and reach the people doing
the most exciting science of the 90s!!

Please call or FAX Teresa Tiganis, Advertising Manager for further details.
Tel. (516) 367-8351, FAX (516) 367-8532.

Editorial office: Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, New York 11724-2203. Phone 516-367-8492; FAX 516-367-8334.

GENOME RESEARCH (ISSN 1054-9803) is published monthly for \$495 (U.S. institutional; \$545 R.O.W.), \$95 (individual making personal payment; \$145 R.O.W., includes airlift) by Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, New York 11724. Periodicals class postage pending is paid at Cold Spring Harbor and additional mailing offices. POSTMASTER: Send address changes to Cold Spring Harbor Laboratory, 10 Skyline Drive, Plainview, New York 11803-2500.

Manager. Personal: U.S. \$95; R.O.W. \$145 (includes airlift). Institutional: U.S. \$495; R.O.W. \$545 (includes airlift). Orders may be sent to Cold Spring Harbor Laboratory Press, Fulfillment Department, 10 Skyline Drive, Plainview, New York 11803-2500. Telephone: Continental U.S. and Canada 1-800-843-4388; all other locations 516-349-1930. FAX 516-349-1946. Personal subscriptions must be prepaid by personal check, credit card, or money order. Claims for missing issues must be received within 4 months of issue date.

Advertising: Teresa Tiganis, Advertising Manager, Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, New York 11724-2203. Phone: 516-

Copyright information: Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$5.00 per copy is paid directly to CCC, 21 Congress Street, Salem, Massachusetts 01970 (1054-9803/96 - \$5.00). This consent does not extend to other kinds of copying, such as copying for general distribution for advertising or promotional purposes, for creating new collective works, or for resale.

Copyright © 1996 by Cold Spring Harbor

GENOME RESEARCH

Volume 6 Number 7
July 1996

RESEARCH PAPERS

- | | | |
|--|--|-----|
| Gene Transfer into Corn Earworm
(<i>Helicoverpa zea</i>) Embryos | James D. DeVault, Keith J. Hughes,
Roger A. Leopold, Odell A. Johnson,
and Sudhir K. Narang | 571 |
| Identity-by-descent Mapping of Recessive Traits
in Livestock: Application to Map the Bovine
<i>Syndactyly</i> Locus to Chromosome 15 | Carole Charlier, Frédéric Farnir,
Paulette Berzi,
Pascal Vanmanshoven,
Benoît Brouwers, Hans Vromans,
and Michel Georges | 580 |
| Sequencing the 500-kb GC-rich Symbiotic
Replicon of <i>Rhizobium</i> sp. NGR234 Using Dye
Terminators and a Thermostable "Sequenase":
A Beginning | Christoph Freiberg, Xavier Perret,
William J. Broughton, and
André Rosenthal | 590 |
| Worldwide Distribution of Human
Y-chromosome Haplotypes | Fabrizio R. Santos,
Néstor O. Bianchi, and
Sérgio D.J. Pena | 601 |
| Bacterial Artificial Chromosome Cloning and
Mapping of a 630-kb Human
Extrachromosomal Structure | Min Wang, Stephanie Shouse,
Barbara Lipes, Ung-Jin Kim,
Hiroaki Shizuya, and Eric Lai | 612 |

LETTERS

- | | | |
|--|--|-----|
| The Genomic Structure of Discoidin Receptor
Tyrosine Kinase | Martin P. Playford, Robin J. Butler,
Xiao Cun Wang, Roy M. Katso,
Inez E. Cooke, and
Trivadi S. Ganesan | 620 |
| A Contiguous High-resolution Radiation Hybrid
Map of 44 Loci from the Distal Portion of the
Long Arm of Human Chromosome 5 | Janet A. Warrington and
John J. Wasmuth | 628 |
-

GENOME METHODS

Uniform Amplification of a Mixture of Deoxyribonucleic Acids with Varying GC Content	Namadev Baskaran, Rajendra P. Kandpal, Ajay K. Bhargava, Michael W. Glynn, Allen Bale, and Sherman M. Weissmann	633
A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization	Dari Shalon, Stephen J. Smith, and Patrick O. Brown	639
Microsatellite Hybrid Capture Technique for Simultaneous Isolation of Various STR Markers	Michal Prochazka	646
Erratum		650

Product News	651
--------------	-----

COVER DNA microarrays for analyzing complex DNA samples. Shown is a two-color fluorescent scan of an 1.8-cm × 1.8-cm yeast array of λ clones of yeast genomic DNA. (For details, see Shalon et al., p. 639.)

A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization

Dari Shalon,^{1,4} Stephen J. Smith,³ and Patrick O. Brown^{1,2,5}

¹Howard Hughes Medical Institute and Departments of ²Biochemistry and ³Molecular and Cellular Physiology, Stanford University, Stanford, California 94305

Detecting and determining the relative abundance of diverse individual sequences in complex DNA samples is a recurring experimental challenge in analyzing genomes. We describe a general experimental approach to this problem, using microscopic arrays of DNA fragments on glass substrates for differential hybridization analysis of fluorescently labeled DNA samples. To test the system, 864 physically mapped λ clones of yeast genomic DNA, together representing >75% of the yeast genome, were arranged into 1.8-cm \times 1.8-cm arrays, each containing a total of 1744 elements. The microarrays were characterized by simultaneous hybridization of two different sets of isolated yeast chromosomes labeled with two different fluorophores. A laser fluorescent scanner was used to detect the hybridization signals from the two fluorophores. The results demonstrate the utility of DNA microarrays in the analysis of complex DNA samples. This system should find numerous applications in genome-wide genetic mapping, physical mapping, and gene expression studies.

Many problems in genome analysis depend on determining what specific sequences are represented in a complex DNA or RNA sample and at what abundance, for example, what genes are represented in a specific chromosome band or YAC clone, what intervals are amplified or deleted in a particular cancer cell, or what genes are expressed in specific cells under specific conditions. As a general approach to this problem, we have developed a system for making microarrays of DNA samples on glass substrates, probing them by hybridization with complex fluorescent-labeled probes, and using a laser-scanning microscope to detect the fluorescent signals representing hybridization. Fluorescent labeling allows for simultaneous hybridization and separate detection of the hybridization signal from two or more probes. This in turn allows very accurate and reliable measurement of the relative abundance of specific sequences in two complex samples.

RESULTS

Array Hybridization Pattern

Figure 1 shows the two-color fluorescent scan of a yeast genomic array following hybridization

with a mixed probe consisting of lissamine-labeled DNA from the 6 largest yeast chromosomes together with fluorescein-labeled DNA from the 10 smallest yeast chromosomes. A red color indicates that yeast sequences present in the lissamine-labeled hybridization probe hybridized to an array element. A yellow-green color indicates that yeast sequences present in the fluorescein-labeled hybridization probe hybridized to an array element. An orange color indicates cross-hybridization of both chromosome pools to an array element (e.g., dispersed repetitive elements, such as Ty1 elements).

Each clone was spotted twice, resulting in duplicate hybridization patterns in adjacent quadrants of the array. Control DNA spots, which were randomly amplified in the same manner as the λ clone array elements, are located in the bottom corner of each quadrant. "A" points to a pair of spots containing total yeast genomic DNA. These spots appear orange because both chromosome pools hybridized to yeast genomic DNA. The negative controls are as follows: "B" points to a pair of spots of wild-type λ DNA, "C" points to a pair of human genomic DNA spots, and "D" points to a pair of ϕ X174 DNA spots. The lack of a hybridization signal at these three negative control spots indicates that the hybridization was specific for yeast sequences.

⁴Present address: Syntent, Inc., Palo Alto, California 94305.

⁵Corresponding author.

E-MAIL pbrown cmgm.stanford.edu, <http://cmgm.stanford.edu/pbrown>; FAX (415) 723-1399.

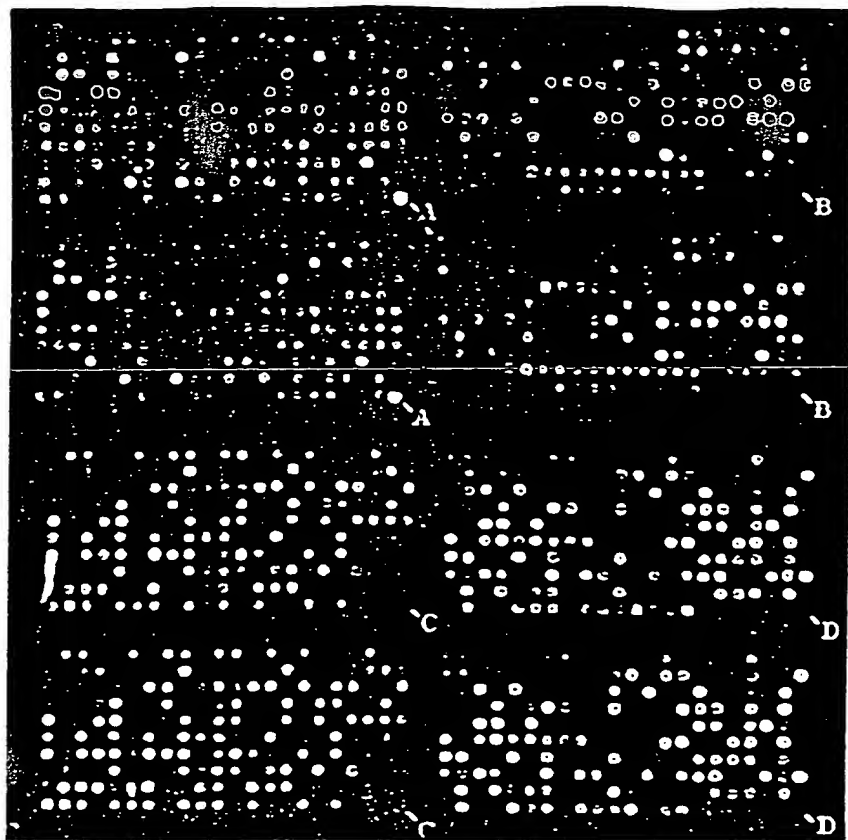


Figure 1 Two-color fluorescent scan of a 1.8-cm \times 1.8-cm yeast array of λ clones of yeast genomic DNA. The DNA spots are spaced at a distance of 380 μ m from center to center. A probe mixture consisting of DNA from the 6 largest yeast chromosomes (4, 7, 12, 13, 15, 16) labeled with lissamine (red dots) and DNA from the 10 smallest yeast chromosomes (1, 2, 3, 5, 6, 8, 9, 10, 11, 14) labeled with fluorescein (yellow-green dots) was hybridized to the array. A pair of yeast genomic DNA spots (A) served as a positive control. The three negative controls are λ DNA (B), human genomic DNA (C), and ϕ X174 DNA (D).

Karyotype Depiction of the Array Hybridization Pattern

The inserts contained in the arrayed λ clones have been mapped physically (Riles et al. 1993). The clones are arrayed in a random but known order on the array. Therefore, using the identity of each clone along with its physical map information, the pattern of hybridization to the yeast array can be represented in the form of a karyotype of the yeast genome, as shown in Figure 2. The color of any segment of the ideogram representing an individual chromosome on the karyotype is directly determined by the ratio of red and green hybridization signals at the array positions of the corresponding clones. The lengths of the discrete colored segments of each chromosome correspond to the physical lengths of the yeast

inserts. The chromosome segments colored black represent either intervals of the genome that are not represented by clones in the library (90%) or false-negative hybridization signals on the array (10%). Most of these false negatives are attributable to failures of the PCR amplification of the λ clones, though occasional failures of the arraying process or nonuniform surface preparation could account for a small fraction of the false-negative signals. The large gap on chromosome 12 is the region coding for ribosomal DNA that was not represented among the arrayed clones. Genomic intervals represented by overlapping clones were assigned a color based on the hybridization signals of only one of the overlapping clones, chosen at random.

Note that in this representation of a yeast karyotype, the largest six chromosomes are mainly colored red. This indicates that most of the arrayed clones that were mapped previously to these six large chromosomes hybridized primarily to the lissamine-labeled probe prepared from the corresponding purified chromosomes. Conversely, the smallest 10 chromosomes are mainly colored green in this image, matching the original CHEF gel isolation of the chromosomes used as the hybridization probe. The experiment was repeated with the yeast genome split into six discrete chromosome pools containing 2–4 chromosomes per pool using CHEF gel electrophoresis. The chromosomes in each pool were extracted from the gel, amplified, and fluorescently labeled. The six chromosome pools were hybridized to six separate yeast arrays. Forty-four λ clones gave a positive hybridization signal on all six arrays indicating that they contain yeast repetitive sequences (data not shown). These 44 clones and 10 clones with very weak hybridization signals were not included in the data set used to produce this karyotype.

There were ~40 anomalous clones, which appear in this karyotype representation as green bands on the chromosome ideogram.

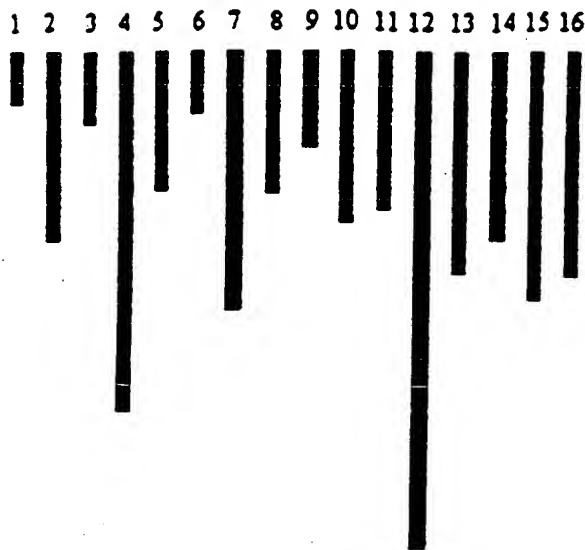


Figure 2 Computer-generated ideogram representing a karyotype of *S. cerevisiae*, based on the normalized hybridization signals from the array shown in Fig. 1. Note that the 6 largest chromosomes are mainly red and the 10 smallest chromosomes are mainly green. Black stripes represent intervals not represented by clones in the array or for which the corresponding clones gave false-negative hybridization signals.

bands on the otherwise green chromosomes. Four randomly chosen examples of these anomalous clones were analyzed by hybridizing the clones to vertical strips cut from a Southern blot of CHEF gel-separated yeast chromosomes. In each case, the hybridization patterns of the anomalous clones corroborated the chromosomal locations assigned by the microarray hybridization results (data not shown). Two clones that were thought to map to the 10 smallest chromosomes were found to hybridize preferentially to the probe representing the 6 largest chromosomes and thus appear as anomalous red bands on the karyotype. Both hybridized to one of the six largest chromosomes on the Southern blot. Similarly, two clones that appear as anomalous green bands on the karyotype were found to hybridize to one of the 10 smallest chromosomes on the Southern blot. Thus, the anomalous clones are probably the result of sample tracking errors or, possibly, of errors in the published restriction-digest-based physical map on which the karyotype representation was based (Riles et al. 1993).

DISCUSSION

The DNA microarray hybridization system reported here is conceptually and functionally

similar to fluorescent in situ hybridization (FISH) to metaphase chromosomes, with three important differences. First, the target elements of the microarrays can, in principle, be any length or composition, from megabase YAC clones or microdissected chromosome bands to individual cDNA clones, to short oligonucleotides. This versatility allows the user to choose characteristics, such as the mapping resolution and genetic complexity of each array element, to suit a particular application. Second, the hybridization signals are localized to discrete elements of known size and location, making them easier to identify and quantitate than the hybridization signals from irregularly shaped metaphase spreads. Third, microarrays are more consistent and potentially amenable to automated production, hybridization, and data analysis than metaphase spreads.

Arrays of DNA samples on porous membranes, for example, dot blots, have long been used as a basic tool in molecular biology. Dot-blot membranes are usually at least 8×12 cm in size, require the use of milliliter volumes of hybridization solution, and are limited, owing to autofluorescence and scattering, to radioactive, chemiluminescent, and colorimetric hybridization detection methods (Ross et al. 1992). Microarrays made on glass surfaces, on the other hand, can be mass-produced and are comparatively inexpensive, convenient, and compatible with fluorescent hybridization detection methods. Furthermore, a glass surface, when appropriately treated, has very low nonspecific binding of labeled hybridization probes, resulting in lower backgrounds than are encountered typically with porous membranes. For hybridizations with very complex probes, the concentration of the labeled probe DNA is a limiting factor in the sensitivity of the assay. Minimizing the volume of the probe solution in a hybridization, by restricting the target to a small area and by using a nonporous substrate, makes it practical to achieve very high probe concentrations.

One important advantage of fluorescently labeled probes is that, unlike most radioactive and chemiluminescent signals, fluorescent signals do not disperse and therefore allow for very dense array spacing. A unique, and probably the most important, advantage of fluorescent probes is that the hybridization signals from two or more differently labeled probes hybridized to the same target element can be detected separately. In this way, two-color hybridization detection allows for a direct and quantitative comparison of the

abundance of specific sequences between two probe mixtures that are hybridized competitively to a single array. The absolute intensity of a hybridization signal at a particular element in an array can vary owing to experimental factors such as variations in the amount of DNA deposited on the array, variations in the hybridization or wash conditions between experiments, or variations in the hybridization characteristics of the different DNA sequences on the array. The ratio of the two signals at any element in an array, however, is relatively insensitive to these confounding factors because they affect both probe mixtures equivalently. This ratio therefore accurately reflects the relative abundance of the cognate sequence in the two probe samples. This is the principle underlying the technique of comparative genomic hybridization (CGH), which is used to detect changes in the copy number of specific chromosomes or chromosomal regions (Kallioniemi et al. 1992). CGH is based on measuring the relative fluorescent hybridization intensities of two genomic-complexity hybridization probes, for example, probes representing genomic DNA from normal and affected tissue samples, which are labeled with two distinct fluorophores and hybridized simultaneously to a metaphase spread. DNA microarray representations of the human genome may provide a more convenient and higher resolution alternative to metaphase chromosomes for CGH.

Cross-hybridization between related sequences is an important problem faced by any hybridization-based assay, including the DNA microarray assay described here. Studies are now in progress to quantitate the extent of cross-hybridization between related sequences of varying homology and length, in DNA microarray hybridizations. The stringency of hybridization and washing can be controlled by varying the salt concentration and temperature as in conventional membrane-based hybridizations. Cross-hybridization caused by repetitive sequences can be minimized by prehybridization of the probe or array with vast excess of unlabeled copies of the repetitive sequences.

Alternative methods have been described for making microarrays of very short DNA sequences, involving photolithography (Pease et al. 1994) or physical masking (Maskos and Southern 1992) methods. These in situ synthesis methods are inherently limited to low complexity array elements consisting of oligonucleotides. For complex probe hybridizations, the efficiency of

hybridization is improved by using DNA fragments substantially longer than oligonucleotides. Moreover, the in situ synthesis approaches to array fabrication depend on prior knowledge of the sequence to be recognized by each array element. The approach described here makes microarrays by transferring tiny volumes of DNA samples from microwell storage plates to a solid substrate. Thus, nucleic acids (or other molecules) of virtually any length or any origin can be arrayed, and knowledge of their sequences is not required.

The arrays used in these experiments do not represent the maximal achievable density of elements. We have found that the spacing between the spots can be decreased by shrinking the contact area of the printing tip and by increasing the hydrophobicity of the glass surface. Microarrays with 100- μm feature size have been tested successfully in pilot experiments (data not shown). Assuming the projected availability of the appropriate physically mapped human genomic clones (Hudson et al. 1995), arrays at 100- μm spacing would allow for 10,000 discrete intervals of the human genome to be represented in a 1- cm^2 array. Such an array could be used for mapping at a resolution of <0.5 Mb. Experiments are in progress to explore the feasibility of such arrays.

Our initial motivation for developing these microarrays arose from the need for abundant and inexpensive genomic arrays for genomic mismatch scanning (GMS) (Nelson et al. 1993), a method of genetic linkage analysis based on identification of the regions of "identity by descent" between affected relative pairs using a single complex-probe hybridization to an array of genomic clones. Experiments using these arrays to map quantitative trait loci in yeast by GMS are currently in progress (J. deRisi, D. Lashkari, L. Penland, L. McAllister, J. McCusker, R. Davis, and P.O. Brown, unpubl.).

Microarrays of cDNA clones, prepared using the system described here, have been used for quantitative monitoring of gene expression patterns in *Arabidopsis* (Schena et al. 1995), *S. cerevisiae* (D. Lashkari, J. deRisi, L. Penland, P.O. Brown, and R. Davis, unpubl.), and human tissues (J. deRisi, M. Bittner, P. Meltzer, L. Penland, J. Trent, and P.O. Brown, unpubl.). We anticipate that DNA microarrays of the kind described here will be useful in additional applications for which conventional dot blots, high-density gridded arrays on porous membranes, or

DNA MICROARRAYS FOR ANALYZING COMPLEX DNA SAMPLES

tions include comparative genomic hybridization (Kallioniemi et al. 1992), sequencing by hybridization (Drmanac et al. 1993), physical mapping of cloned or amplified sequences (Billings et al. 1991), and economical distribution of reagents for integrated genetic and physical mapping based on a common set of arrayed clones (Zehetner and Lehrach 1994).

METHODS

Amplification of Target DNA Elements

The array elements were prepared from physically mapped λ clones (Riles et al. 1993). The λ clones were amplified using randomly primed polymerase chain reaction (PCR) based on published and unpublished protocols (Bohlander et al. 1992; S. Nelson, unpubl.). The phage lysates were amplified in a 10- μ l PCR reaction using 5 μ M final concentration of primer A (GCTATCTTCAAGATCANNNNNN), 200 μ M dNTPs, and 1 unit of *Taq* polymerase. Round A consisted of five cycles at 94°C for 1 min, 25°C for 1.5 min, 25–72°C over 7 min, and 72°C for 3 min using *Taq* polymerase (BMB). For round B, the reaction volume was brought up to 100 μ l for a final concentration of 2 μ M of primer B (GCTATCTTCAAGATCA), 200 μ M dNTPs, and 4 units of *Taq* polymerase. Round B consisted of 30 cycles of 94°C for 1 min, 56°C for 2 min, and 72°C for 3 min. The amplification was performed in 96-well plates using crude phage lysates as the templates, resulting in an amplification of both the 35-kb λ vector and the 5-kb to 15-kb yeast insert sequences as a distribution of PCR products between 250 bp and 1500 bp in length.

The PCR products were purified and transferred into TE (10 mM Tris, 1 mM EDTA at pH 8.0) buffer using Sephadex G50 gel filtration (Pharmacia) and evaporated to dryness at room temperature overnight. Each of the 864 am-

plified λ clones was rehydrated in 15 μ l of 3 \times SSC (20 \times SSC = 3 M NaCl, 0.3 M Na₃ citrate) in preparation for spotting onto the glass under normal room temperature conditions.

Preparation of DNA Microarrays

The microarrays were fabricated on poly-L-lysine coated microscope slides (Sigma). A custom-built arraying machine, consisting of four tweezer-like printing tips mounted 9 mm apart on a computer-controlled robotic stage (Shalon 1996), loaded 1 μ l of the concentrated PCR product directly from corresponding clusters of four wells of 96-well storage plates and deposited ~5 nl of each sample onto each of 40 slides. Surface tension loaded the sample into the printing tip directly from the microwell plate and held the sample in the tip during the printing operation. Printing was achieved by lightly tapping the tip against the glass surface. The open-capillary design allowed for rapid rinsing and drying of the tips between samples. Figure 3 shows the layout of the arraying machine. Figure 4 shows a detailed view of the four printing tips and the staggered printing pattern on the microscope slides. Adjacent samples were spotted 380 μ m apart on the slides. After each set of four samples was printed onto 40 slides, the printing tips were rinsed with a jet of water for 2 sec and then dried by lowering the tips onto a sponge for 2 sec. The process was repeated for all 864 samples and eight control spots.

After the spotting operation was complete, the slides were rehydrated in a humid chamber at room temperature for 2 hr, baked in an 80°C vacuum oven for 2 hr, then rinsed in 0.1% sodium dodecyl sulfate (SDS) to remove unadsorbed DNA. To reduce nonspecific adsorption of the labeled hybridization probe to the poly-L-lysine coated glass surface, the slides were treated with succinic anhydride. One gram of succinic anhydride was dissolved in 100 ml of 1-methyl-2-pyrrolidinone and then 100 ml of 0.2 M boric acid (pH 8.0) was added. The arrays were soaked in this solution for 10 min and then rinsed in distilled water four times for 5 min each. Immediately before use, the arrayed DNA elements were denatured by placing the slide in distilled water at 90°C for 2 min.

Amplification and Labeling of Hybridization Probe

The 16 chromosomes of *Saccharomyces cerevisiae* were separated using a contour-clamped homogeneous electric field (CHEF) agarose gel apparatus (Bio-Rad) (Chu et al. 1986). The 6 largest chromosomes were isolated in one gel slice and the smallest ten chromosomes in a second gel slice. The DNA from each slice was recovered using a gel extraction kit

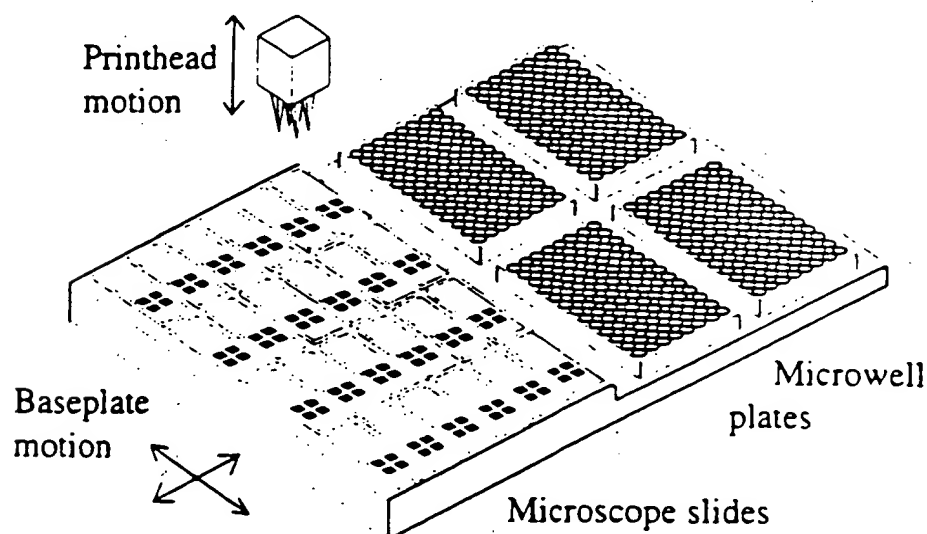


Figure 3 The layout of the arraying machine. All motions are under computer control. For more details of the arraying machine, see web page <http://cmgm.stanford.edu/pbrown>.

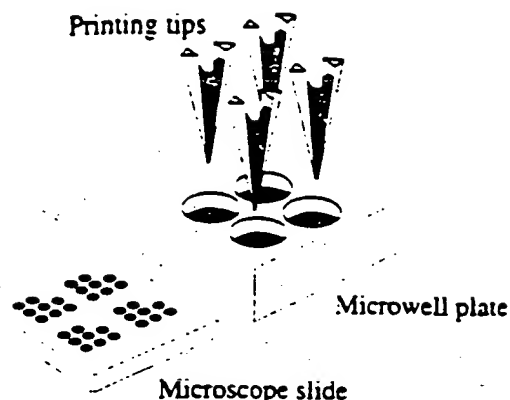


Figure 4 A close-up view of the four open-capillary printing tips. The tips are 9 mm apart and fit into four adjacent wells of a standard microwell plate and print arrays in a staggered fashion on microscope slides. For more details of the printing tips, see web page <http://cmgm.stanford.edu/pbrown>.

(Qiagen) and randomly amplified in a manner similar to that used in amplifying the target λ clones (Grothues et al. 1993). The main difference between this amplification procedure and the one used for the λ array elements is a filtration step between rounds A and B to remove primers and the use of a random 9-mer 3' end on primer A. Following amplification, 2.5 μ g of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham) with a lissamine-conjugated nucleotide analog (DuPont NEN) for the pool containing the 6 largest chromosomes and with a fluorescein-conjugated nucleotide analog (BMB) for the pool containing the smallest 10 chromosomes. The two fluorescent-labeled pools were mixed and concentrated using an ultrafiltration device (Amicon).

Hybridization

Five micrograms of the hybridization probe, consisting of both chromosome pools in 7.5 μ l of TE, was denatured in a boiling water bath and then snap-cooled on ice. Concentrated hybridization solution (2.5 μ l) was added to a final concentration of $5 \times$ SSC/0.1% SDS. The entire 10 μ l of probe solution was transferred to the array surface, covered with a coverslip, placed in a custom-built single-slide humidity chamber, and incubated in a 60°C water bath for 12 hr. The custom-built waterproof slide chamber has a cavity just slightly bigger than a microscope slide and was kept at 100% humidity internally by the addition of 2 μ l of water in a corner of the chamber. The slide was rinsed in $5 \times$ SSC/0.1% SDS for 5 min and then in $0.2 \times$ SSC/0.1% SDS for 5 min. All rinses were at room temperature. The array was then air dried, and a drop of antifade (Molecular Probes) was applied to the array under a 24-mm \times 30-mm coverslip in preparation for scanning.

Detection and Analysis

A custom-built laser scanner was used to detect the two-

color fluorescence hybridization signals from 1.8-cm \times 1.8-cm arrays at 20- μ m resolution. The glass substrate slide was mounted on a computer-controlled, two-axis translation stage (PM-500, Newport, Irvine, CA) that scanned the array over an upward-facing microscope objective (20 \times , 0.75NA Fluor, Nikon, Melville, NY) in a bi-directional raster pattern. A water-cooled Argon/Krypton laser (Innova 70 Spectrum, Coherent, Palo Alto, CA), operated in multiline mode, allowed for simultaneous specimen illumination at 488.0 nm and 568.2 nm. These two lines were isolated by a 488/568 dual-band excitation filter (Chroma Technology, Brattleboro, VT). An epifluorescence configuration with a dual-band 488/568 primary beam splitter (Chroma) excited both fluorophores simultaneously and directed fluorescence emissions toward the two-channel detector. Emissions were split by a secondary dichroic mirror with a 565 transition wavelength onto two multialkali cathode photomultiplier tubes (PMT; R928, Hamamatsu, Bridgewater, NJ), one with an HQ535/50 bandpass barrier filter and the other with a D630/60 bandpass barrier filter (Chroma). Preamplified PMT signals were read into a personal computer using a 12-bit analog-to-digital conversion board (RTI-834, Analog Devices, Norwood, MA), displayed in a graphics window, and stored to disk for further rendering and analysis. The back aperture of the 20 \times objective was deliberately underfilled by the illuminating laser beam to produce a large-diameter illuminating spot at the specimen (5- μ m to 10- μ m half-width). Stage scanning velocity was 100 mm/sec, and PMT signals were digitized at 100 μ sec intervals. Two successive readings were summed for each pixel, such that pixel spacing in the final image was 20 μ m. Beam power at the specimen was ~ 5 mW for each of the two lines.

The scanned image was despeckled using a graphics program (Hijaak Graphics Suite) and then analyzed using a custom image gridding program that created a spreadsheet of the average red and green hybridization intensities for each spot. The red and green hybridization intensities were corrected for optical cross talk between the fluorescein and lissamine channels, using experimentally determined coefficients.

ACKNOWLEDGMENTS

This research was supported by grant HG00450 from the National Institutes of Health-National Center for Human Genome Research, a National Science Foundation graduate fellowship to D.S., and by the Howard Hughes Medical Institute. P.O.B. is an assistant investigator of the Howard Hughes Medical Institute. We thank John Mulligan and John McCusker for help in preparing and amplifying the λ clones used in the arrays, Ren Xin Xia for writing the scanner control software and the image gridding and automatic karyotyping programs, Jeff van Ness at Darwin Molecular Corporation for suggesting the use of succinic anhydride, Stan Nelson, Linda McAllister, Joe deRisi, and Lolita Penland for helpful suggestions in the course of this work, and Joe deRisi and Linda McAllister for helpful comments on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Billings, P.R., C.L. Smith, and C.R. Cantor. 1991. New techniques for physical mapping of the human genome. *FASEB J.* 5: 28-34.
- Bohlander, S.K., R. Espinosa III, M.M. LeBeau, J.D. Rowley, and M.O. Diaz. 1992. A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 13: 1322-1324.
- Chu, G., D. Vollrath, and R. Davis. 1986. Separation of large DNA molecules by contour clamped homogeneous electric fields. *Science* 234: 1582-1585.
- Drmanac, R., S. Drmanac, Z. Strezoska, T. Paunesku, I. Labat, M. Zeremski, J. Snoddy, W.K. Funkhouser, B. Koop, L. Hood, et al. 1993. DNA sequence determination by hybridization: A strategy for efficient large-scale sequencing. *Science* 260: 1649-1652.
- Gr thues, D., C.R. Cantor, and C.L. Smith. 1993. PCR amplification of megabase DNA with tagged random primers (T-PCR). *Nucleic Acids Res.* 21: 1321-1322.
- Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S.H. Xu, et al. 1995. An STS-based map of the human genome. *Science* 270: 1945-1954.
- Kallioniemi, A., O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, and D. Pinkel. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258: 818-821.
- Maskos, U. and E.M. Southern. 1992. Parallel analysis of oligodeoxynucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation. *Nucleic Acids Res.* 20: 1675-1678.
- Nelson, S.F., J.H. McCusker, M. Sander, Y. Kee, P. Modrich, and P.O. Brown. 1995. Genomic mismatch scanning: A new approach to genetic linkage mapping. *Nature Genet.* 4:11-17.
- Pease, A.C., D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P. Fodor. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci.* 91: 5022-5026.
- Riles, L., J.E. Dutchik, A. Baktha, B.K. McCauley, E.C. Thayer, M.P. Leckie, V.V. Braden, J.E. Depke, and M.V. Olson. 1993. Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics* 134: 81-150.
- Ross, M.T., J.D. Hoheisel, A.P. Monaco, Z. Larin, G. Zehetner, and H. Lehrach. 1992. High density gridded YAC filters: Their potential as genome mapping tools. In *Techniques for the analysis of complex genomes* (ed. Rakesh Anand), pp. 137-153. Academic Press, London, UK.
- Schena, M., D. Shal n, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470.
- Shalon, D.. 1995. "DNA micro arrays: A new tool for genetic analysis." Ph.D. thesis, Stanford University, Stanford, CA.
- Zehetner, G. and H. Lehrach. 1994. The reference library system—Sharing biological material and experimental data. *Nature* 367: 489-491.

Received March 4, 1996; accepted in revised form May 9, 1996.

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

RENU A. HELLER^{*†}, MARK SCHENA^{*}, ANDREW CHAI^{*}, DARI SHALON[‡], TOD BEDLION[‡], JAMES GILMORE[‡], DAVID E. WOOLLEY[§], AND RONALD W. DAVIS^{*}

^{*}Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; [‡]Synteni, Palo Alto, CA 94306; and [§]Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

Contributed by Ronald W. Davis, December 27, 1996

ABSTRACT cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synovocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Gro α and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm². These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix-degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endopeptidases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3' end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF- α , tumor necrosis factor α ; IL, interleukin; TGF- β , transforming growth factor β ; G-CSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.

To whom reprint requests should be sent at the present address: Roche Bioscience, 53–1, 3401 Hillview Avenue, Palo Alto, CA 94304.

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 μ l of 3 \times standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 μ g/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF- α) at 50 ng/ml, interleukin (IL)-1 β at 30 ng/ml, or transforming growth factor- β (TGF- β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 μ l of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 μ l of hybridization buffer (5 \times SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β -actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Genes Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β -actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-

	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 HAT1	HAT1 HAT1	HAT4 HAT4	HAT4 HAT4	HAT22 HAT22	HAT22 HAT22	YES23 YES23	YES23 YES23	BACTIN β -actin	G3PDH G3PDH
B	IL1A IL-1 α	IL1B IL-1 β	IL1RA IL-1RA	IL2 IL-2	IL3 IL-3	IL4 IL-4	IL6 IL-6	IL6R IL-6R	IL7 IL-7	CFOS c-fos	CJUN c-jun	RFRA1 Rat Fra-1
C	IL8 IL-8	IL9 IL-9	IL10 IL-10	ICE ICE	IFNG IFN γ	GCSF G-CSF	MCSF M-CSF	GMCSF GM-CSF	TNFB1 TNF β	CREL c-rel	NFKB50 NF- κ Bp50	NFKB65.1 NF- κ Bp65
D	TNFA1 TNF α	TNFA2 TNF α	TNFA3 TNF α	TNFA4 TNF α	TNFA5 TNF α	TNFR1 TNF α	TNFR2 TNF α	TNFR1 TNF α	TNFR2 TNF α	NFKB65.2 NF- κ Bp65	IKB I κ B	CREB2 CREB2
E	STR1 Strom-1	STR2/3 Strom-2	STR3 Strom-3	COL1 Col-1	COL1/3 Col-1/3	COL2/1 Col-2	COL2/2 Col-2	COL3 Col-3	COX1 Cox-1	COX2 Cox-2	12LO 12-LO	15LO 15-LO
F	GELA1 Gel-A	GELB Gel-B	HME Elastase	MTMMP MT-MMP	PUMP1 Matrilysin	TIMP1 TIMP-1	TIMP2 TIMP-2	TIMP3 TIMP-3	ICAM1 ICAM-1	VCAM VCAM	SLO.1 S-LO	CPLA2.2 cPLA2
G	EGF EGF	FGFA FGF acidic	FGFB FGF basic	IGF1 IGF-1	IGFII IGF-II	TGFA TGF α	TGFB TGF β	PDGFB PDGF β	CALCTN Calctonin	GH1 GH-1	GRO GRO1 α	GCR GCR
H	MCP1.1 MCP-1	MCP1.1 MCP-1	MIP1A MIP-1 α	MIP1B MIP-1 β	MIP MIP	RANTES RANTES	INOS iNOS	LDLR LDLR	ALU.1 IL-10	ALU.2 TNFRp70	ALU.3 IL-10	POLYA LDLR

A. thaliana controls

Human controls

Cytokines and related genes

Transcription factors and related genes

MMP's and related genes

Chemokines

Growth factors and related genes

Other genes

FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 $\mu\text{g}/\mu\text{l}$ or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 $\mu\text{g}/\mu\text{l}$ sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ~ 300 bp, arrayed at a concentration of 1 $\mu\text{g}/\mu\text{l}$. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3' proximal regions, because the signal was reduced if the target fragment was biased toward the 5' end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3' proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metal-

loproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

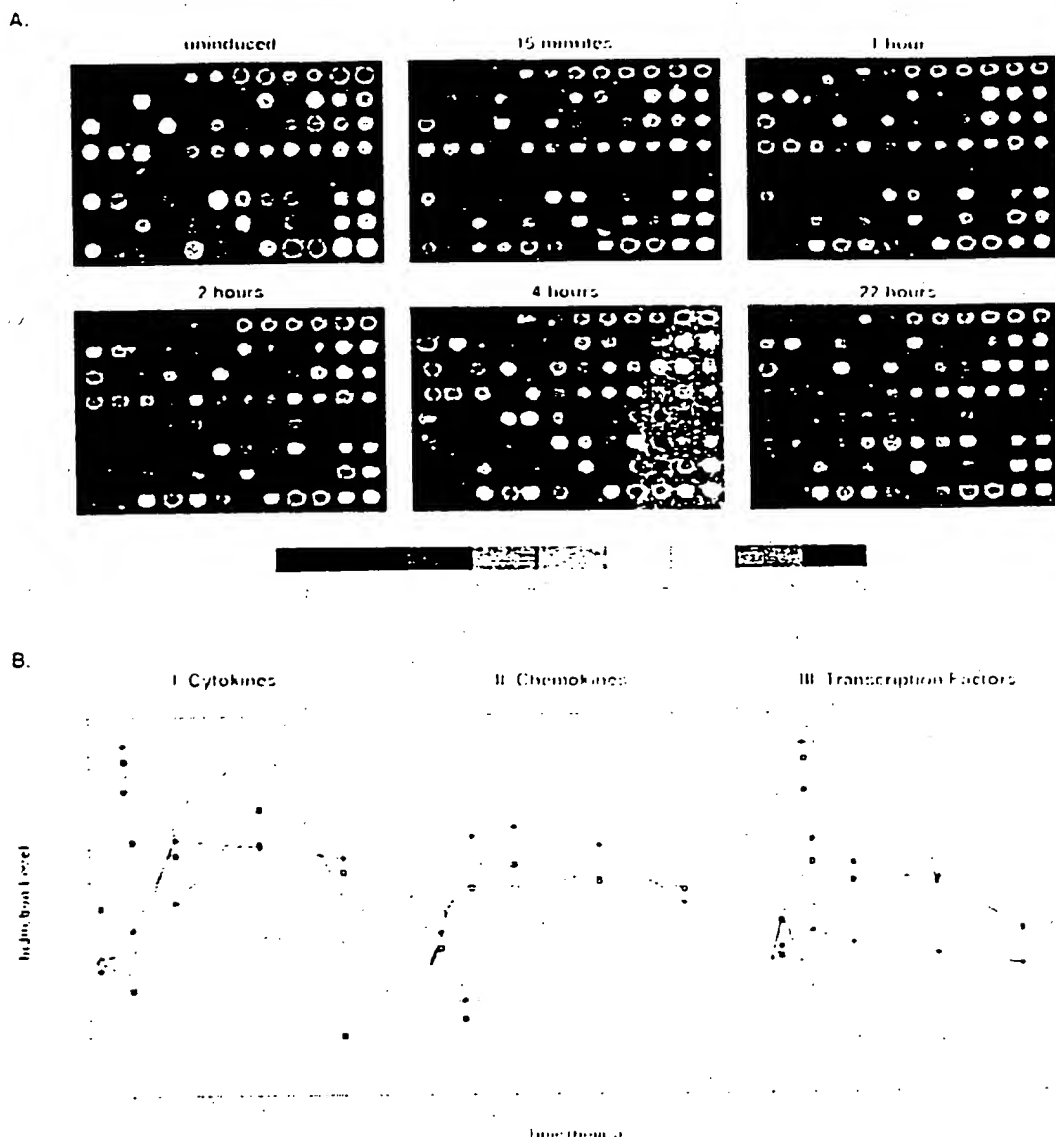


FIG. 2. Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (A) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (B I–III) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatory cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 24). Many other cytokine genes were also transiently activated, such as IL-1 α and - β , IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1 β , more so than MIP-1 α , and Gro α or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were *c-fos*, *fra-1*, *c-jun*, NF- κ Bp50, and I κ B, with *c-rel* expression observed even in the uninduced state (Fig. 2B). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-

type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 34). HME induction was estimated to be ~50-fold and was greater than any of the other MMPs examined (Fig. 3B). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1 β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2B and 3B) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1 β , IL-8, MIP-1 β , TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of *c-jun*, GCSF, IL-3, TNF- β , MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3' end of the gene gave no

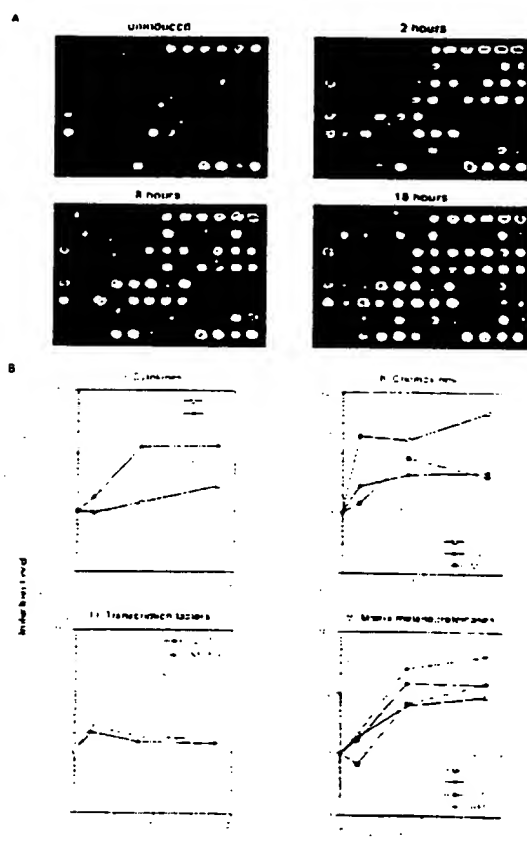


FIG. 3. Time course for IL-1 β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (A) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (B-I) Relative levels of selected genes at different time points compared with time zero.

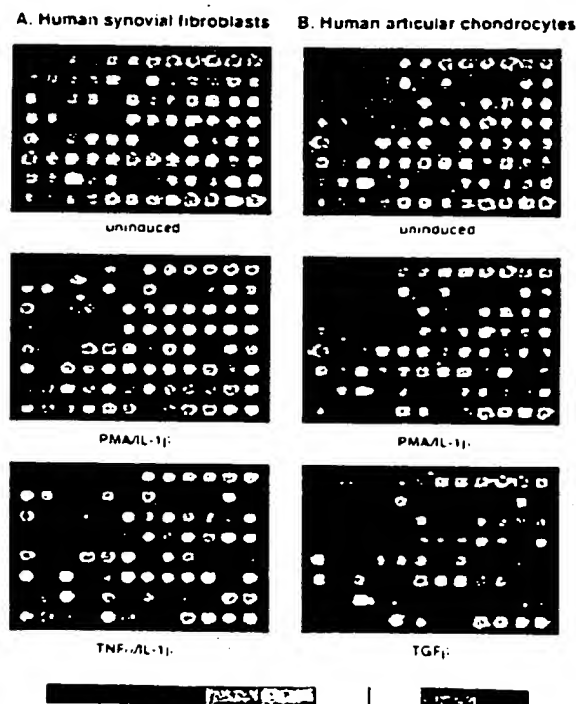


FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1 β , or TNF and IL-1 β , or TGF- β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1 α , and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF- β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in

certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Gro α , and MMP expression was more pronounced than in the others.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1 α , MIP-1 β , IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1 α and IL-1 β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarchy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

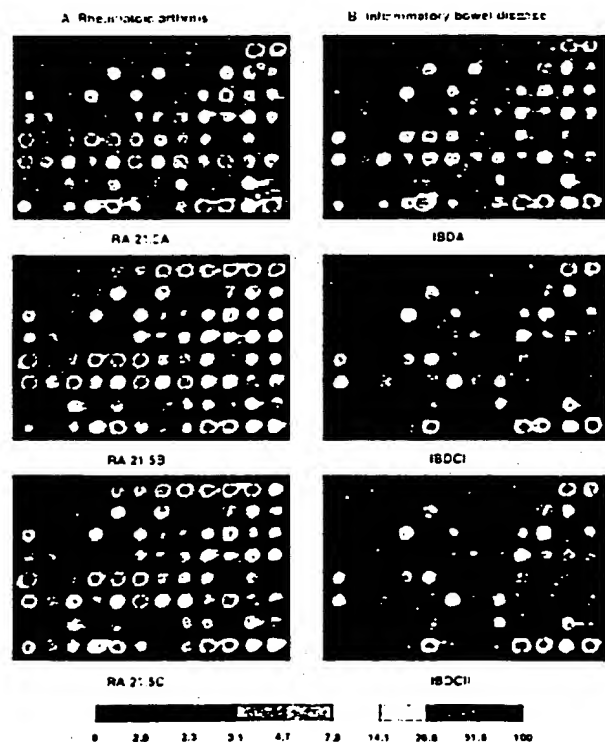


FIG. 5. Expression profiles of RA tissue (A) and IBD tissue (B). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF- β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Gro α , MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Gro α . With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1 β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synovocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Gro α , is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It down-regulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synovocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1 β , the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new

targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

We would like to thank the following individuals for their help in obtaining reagents or providing cDNA clones to use as templates in target preparation: N. Arai, P. Cannon, D. R. Cohen, T. Curran, V. Dixit, D. A. Geller, G. I. Goldberg, M. Karin, M. Lotz, L. Mattisian, G. Nolan, C. Lopez-Otin, T. Schall, S. Shapiro, I. Verma, and H. Van Wart. Support for R.W.D., M.S., and R.A.H. was provided by the National Institutes of Health (Grants R37HG00198 and HG00205).

1. Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995) *Science* 270, 467-470.
2. Shalon, D., Smith, S., & Brown, P. O. (1996) *Genome Res.* 6, 639-645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* 93, 10614-10619.
4. Feldmann, M., Brennan, F. M., & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* 85, 307-310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410-460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger, P. M., & Geng, Y. (1995) *J. Rheumatol.* 22, Supplement 43, 104-108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A., & Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* 4, 197-250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A., & Reithmüller, G. (1988) *Int. J. Cancer* 41, 456-461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P., & Heller, R. A. (1996) *J. Biol. Chem.* 271, 23577-23581.
10. Gadhier, S. J., & Woolley, D. E. (1987) *Rheumatol. Int.* 7, 13-22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* 322, 1277-1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S., & Sledge, C. B. (Saunders, Philadelphia), 5th Ed., pp. 5001-5047.
13. Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K., & Firestein, Gary S. (1991) *J. Immunol.* 146, 3365-3371.
14. Firestein, G. S., Alvaro-Garcia, J. M., & Maki, R. (1990) *J. Immunol.* 144, 3347-3352.
15. Pradines-Figueras, A., & Raetz, C. R. H. (1992) *J. Biol. Chem.* 267, 23261-23268.
16. Shapiro, S. D., Kobayashi, D. L., & Ley, T. J. (1993) *J. Biol. Chem.* 268, 23824-23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J., & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* 93, 3042-3046.
18. Cerretti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K., & Black, R. A. (1992) *Science* 256, 97-100.
19. Miura, M., Zhu, H., Rotello, R., Hartweg, E. A., & Yuan, J. (1993) *Cell* 75, 653-660.
20. Arai, K., Lee, F., Miyajima, A., Shiochiro, M., Arai, N., & Takashi, Y. (1990) *Annu. Rev. Biochem.* 59, 783-836.
21. Geiser, T., Dewald, B., Ehrengruber, M. U., Lewis, L. C., & Baggiolini, M. (1993) *J. Biol. Chem.* 268, 15419-15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A., & Horuk, R. (1993) *J. Biol. Chem.* 268, 1338-1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N., & Fish, E. N. (1995) *Clin. Exp. Immunol.* 101, 398-407.
24. Roesser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A., & Worwood, M. (Academic, New York), Vol. 2, pp. 605-640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V., & Torti, F. M. (1995) *J. Biol. Chem.* 270, 15285-15293.

1

Molecular Cloning

A LABORATORY MANUAL

SECOND EDITION

J. Sambrook

UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER

E.F. Fritsch

GENETICS INSTITUTE

T. Maniatis

HARVARD UNIVERSITY



**Cold Spring Harbor Laboratory Press
1989**

Molecular Cloning

A LABORATORY MANUAL
SECOND EDITION

All rights reserved
© 1989 by Cold Spring Harbor Laboratory Press
Printed in the United States of America

9 8 7 6 5 4 3 2 1

Book and cover design by Emily Harste

Cover: The electron micrograph of bacteriophage λ particles stained with uranyl acetate was digitized and assigned false color by computer. (Thomas R. Broker, Louise T. Chow, and James I. Garrels)

Cataloging in Publications data

Sambrook, Joseph

Molecular cloning: a laboratory manual / E.F.

Fritsch, T. Maniatis—2nd ed.

p. cm.

Bibliography: p.

Includes index.

ISBN 0-87969-309-6

1. Molecular cloning—Laboratory manuals. 2. Eukaryotic cells—Laboratory manuals. I. Fritsch, Edward F. II. Maniatis, Thomas III. Title.

QH442.2.M26 1987

574.87'3224—dc19

87-35464

Researchers using the procedures of this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$0.10 per page is paid directly to CCC, 21 Congress St., Salem MA 01970. [0-87969-309-6/89 \$00 + \$0.10] This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale.

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory, Box 100, Cold Spring Harbor, New York 11724. Phone: 1-800-843-4388. In New York (516)367-8423.

Analysis of RNA

A number of methods have been developed to quantitate, measure the size of, and map the 5' and 3' termini of specific mRNA molecules in preparations of cellular RNA. These include:

- *Northern hybridization (RNA blotting)*, in which the size and amount of specific mRNA molecules in preparations of total or poly(A)⁺ RNA are determined (Alwine et al. 1977, 1979). The RNA is separated according to size by electrophoresis through a denaturing agarose gel and is then transferred to activated cellulose (Alwine et al. 1977; Seed 1982b), nitrocellulose (Goldberg 1980; Thomas 1980; Seed 1982a), or glass or nylon membranes (Bresser and Gillespie 1983) (see below). The RNA of interest is then located by hybridization with radiolabeled DNA or RNA followed by autoradiography.
- *Dot and slot hybridization*, in which an excess of radiolabeled probe is hybridized to RNA that has been immobilized on a solid support (Kafatos et al. 1979; Thomas 1980; White and Bancroft 1982). Densitometric tracings of the resulting autoradiographs can allow comparative estimates of the amount of the target sequence in various preparations of RNA.
- *Mapping RNA using nuclease S1 or ribonuclease*, in which the precise positions of the 5' and 3' termini of the mRNA and the locations of splice junctions can be rigorously determined (Berk and Sharp 1977; Weaver and Weissmann 1979). Labeled or unlabeled RNA or DNA probes derived from various segments of the genomic DNA are hybridized to mRNA, often under conditions favoring the formation of DNA:RNA hybrids (Casey and Davidson 1977). The products of the hybridization are then digested with nuclease S1 or RNAase under conditions favoring digestion of single-stranded nucleic acids only. Analysis of the digestion products by gel electrophoresis yields important quantitative and qualitative information about the mRNA structure.
- *Primer extension*, in which a small radiolabeled fragment of DNA is hybridized to the mRNA and used as a primer for reverse transcriptase. The resulting product should extend to the extreme 5' terminus of the mRNA, and thus the size of the product reflects the number of nucleotides from the position of the label to the 5' terminus of the mRNA.
- *Solution hybridization*, in which the absolute concentration of the sequence of interest is calculated from the rate of hybridization of a small amount of a specific radioactive probe with a known quantity of purified cellular RNA (see, e.g., Roop et al. 1978; Durnam and Palmiter 1983). Alternatively, an excess of a radiolabeled probe is incubated with a known amount of RNA. The concentration of the RNA of interest can then be estimated from the amount of radioactivity that becomes resistant to nuclease S1 (see, e.g., Favalaro et al. 1980; Beach and Palmiter 1981; Williams et al. 1986).

- *Filter hybridization*, in which purified cellular RNA is end-labeled with ^{32}P and hybridized to a large excess of the homologous DNA that has been immobilized on a solid support (Williams et al. 1986).

Below we describe northern hybridization. Dot and slot hybridization of both crude and purified preparations of RNA are described beginning on page 7.53; nuclease-S1 and RNAase analysis of specific hybrids, beginning on pages 7.58 and 7.71, respectively; and analysis of mRNA by primer extension, beginning on page 7.79.

Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes

(Human Genome Project/DNA chip/gene discovery/T cell)

MARK SCHENA^{*†}, DARI SHALON[‡], RENU HELLER^{*}, ANDREW CHAI^{*}, PATRICK O. BROWN[§], AND RONALD W. DAVIS^{*}

^{*}Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; [†]Systems, Palo Alto, CA 94304; [‡]Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305

Contributed by Ronald W. Davis, June 26, 1996

ABSTRACT Microarrays containing 1046 human cDNAs of unknown sequence were printed on glass with high-speed robotics. These 1.0-cm² DNA "chips" were used to quantitatively monitor differential expression of the cognate human genes using a highly sensitive two-color hybridization assay. Array elements that displayed differential expression patterns under given experimental conditions were characterized by sequencing. The identification of known and novel heat shock and phorbol ester-regulated genes in human T cells demonstrates the sensitivity of the assay. Parallel gene analysis with microarrays provides a rapid and efficient method for large-scale human gene discovery.

Biology has entered the genome era (1). Complete genome sequences for all of the model organisms and human will probably be available by the year 2003 (2). Torrents of human expressed sequence tags (ESTs) provide a starting point for elucidating the function of tens of thousands of cognate genes (3). Genome analysis will provide insights into growth, development, differentiation, homeostasis, aging, and the onset of diseases (1–3). A detailed understanding of the human genome will require the implementation of sophisticated methods for gene expression analysis and gene discovery.

Recently, a microarray-based method for high-throughput monitoring of plant gene expression was described (4). This "chip"-based approach involved using microarrays of cDNA clones as gene-specific hybridization targets to quantitatively measure expression of the corresponding plant genes (4, 5). A two-color fluorescence labeling and detection scheme facilitated sensitive differential expression analysis of different plant tissues (4, 5). The efficiency of this approach for studies in higher plants suggested the use of this method for human genome analysis (4–7). Here, we report the use of cDNA microarrays for human gene expression monitoring, biological investigation, and gene discovery.

MATERIALS AND METHODS

Human cDNA Clones. The cDNA library was made with mRNA from human peripheral blood lymphocytes transformed with the Epstein-Barr virus. Inserts >600 bp were cloned into the lambda vector λ YES-R to generate 10⁵–10⁶ recombinants. Bacterial transformants were obtained by infecting *E. coli* strain JM107/ΔKC. Colonies were picked at random and propagated in a 96-well format, and miniprep DNA was prepared by alkaline lysis using REAL preps (Qiagen, Chatsworth, CA). Inserts were amplified by PCR in a 96-well format using primers (PAN132, 5'-CCTC-TATACTTTAACGTCAGG; and PAN133, 5'-TTGTGTG-GAATTGTGAGCGG) complementary to the λ YES polylinker and containing a six-carbon amino modification

(Glen Research, Sterling, VA) on the 5' end. PCR products were purified in a 96-well format using QIAquick columns (Qiagen).

Microarray Preparation. Amino-modified PCR products were suspended at a concentration of 0.5 mg/ml in 3× standard saline citrate (SSC) and arrayed from 96-well microtiter plates onto silylated microscope slides (CEL Associates, Houston) using high-speed robotics (4–7). A total of 1056 cDNAs, representing 1046 human clones and 10 *Arabidopsis* controls, were arrayed in 1.0-cm² areas. Printed arrays were incubated for 4 hr in a humid chamber to allow rehydration of the array elements and rinsed, once in 0.2% SDS for 1 min, twice in H₂O for 1 min, and once for 5 min in sodium borohydride solution (1.0 g of NaBH₄ dissolved in 300 ml of PBS and 100 ml of 100% ethanol). The arrays were submerged in H₂O for 2 min at 95°C, transferred quickly into 0.2% SDS for 1 min, rinsed twice in H₂O, air dried, and stored in the dark at 25°C.

Fluorescent Probes. Tissue mRNAs were purchased (CLONTECH). Jurkat mRNA was isolated as described by Schena *et al.* (4). Probes were made as described (4) with several modifications. The reverse transcriptase used here was Superscript II RNase H⁻ (GIBCO). The Cy5-dCTP was purchased from Amersham. Each reverse transcription reaction contained 3.0 μg of total human mRNA. *Arabidopsis* control mRNAs were made by *in vitro* transcription of cloned HAT4, HAT22, and YesA1-23 cDNAs (4, 8, 9) using an RNA Transcription Kit (Stratagene). For quantitation, the mRNAs were doped into the reverse transcription reaction at ratios of 1:100,000, 1:10,000, and 1:1000 (wt/wt) respectively. Following the reverse transcription step, samples were treated with 2.5 μl of 1 M sodium hydroxide for 10 min at 37°C, then neutralized by adding 2.5 μl of 1 M Tris-HCl (pH 6.8) and 2.0 μl of 1 M HCl. Probe mixtures contained cDNA products derived from 3 μg of total mRNA, suspended in 5.0 μl of hybridization buffer (5× SSC plus 0.2% SDS).

Hybridization and Scanning. Probes were hybridized to 1.0-cm² microarrays under a 14 × 14 mm glass coverslip for 6–12 hr at 60°C in a custom-built hybridization chamber (4–7). Arrays were washed for 5 min at room temperature (25°C) in low stringency wash buffer (1× SSC/0.2% SDS), then for 10 min at room temperature in high stringency wash buffer (0.1× SSC/0.2% SDS). Arrays were scanned in 0.1× SSC using a fluorescence laser scanning device (4–7), fitted with a custom filter set (Chroma Technology, Brattleboro, VT). Accurate differential expression measurements (i.e., final fluorescence ratios) were obtained by taking the average of the ratios of two independent hybridizations.

Abbreviation: EST, expressed sequence tag.

Data deposition: The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U56654–U56660).

[†]To whom reprint requests should be addressed. e-mail: schena@cmgm.stanford.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Cell Culture. Jurkat cells were grown in a tissue culture incubator (37°C and 5% CO₂) in RPMI medium supplemented with 10% fetal bovine serum, 100 µg of streptomycin per ml, and 500 units of penicillin per ml. Heat shock corresponded to a 4-hr incubation at 43°C. Phorbol ester treated cells were grown for 4 hr in the presence of 50 ng of phorbol 12-myristate 13-acetate (PMA) per ml.

RNA Blotting. Dot blots were performed as described (4).

DNA Sequencing. Sequences were obtained using the PAN132 and PAN133 primers and a 373A automated sequencer, according to the instructions of the manufacturer (Applied Biosystems).

Computer Graphics and Informatics. Pseudocolor representations of fluorescent images were made with National Institutes of Health IMAGE software (version 1.52). Software for differential expression representations was purchased from Imaging Research (St. Catherine's, ON, Canada). Sequence searches were made to the nonredundant nucleotide data base at the National Center for Biotechnology Information (NCBI) using Macintosh BLAST software. The EST data base was accessed via the World Wide Web (<http://www.ncbi.nlm.nih.gov/>).

RESULTS

Gene Discovery and the Heat Shock Response. Microarrays were used to examine the heat shock response in cultured human T (Jurkat) cells. Control (37°C) and heat-treated (43°C) cells were harvested and lysed, and total mRNA from the two cell samples was labeled by reverse transcriptase incorporation of fluorescein- and Cy5-dCTP, respectively. In a second set of labeling reactions, the fluorescent groups were "swapped" such that samples from control and heat-treated

samples were labeled with Cy5- and fluorescein-dCTP, respectively. Each pair of fluorescent probes was hybridized to a 1056-element microarray. The arrays were washed at high stringency and scanned with a confocal laser scanning device to detect emission of the two fluorescent groups.

Hybridization signals were observed to >95% of the human cDNA array elements, but not to any of the *Arabidopsis* negative controls (Fig. 1). Fluorescence intensities spanned more than three orders of magnitude for the 1046 array elements surveyed (Fig. 1). Comparative expression analysis of heat shocked versus control cells in the two experiments revealed 17 array elements that displayed altered fluorescence ratios of ≥ 2.0 -fold (Figs. 1 and 24). Of the 17 putative differentially expressed genes, 11 were induced by heat shock treatment and 6 displayed modest repression (Figs. 1 and 2.4).

To determine the identity of the heat-regulated genes, cDNAs corresponding to each of the 17 array elements were sequenced on the proximal and distal end. Data base searches revealed perfect matches for 14 of the 17 clones, and in each case proximal and distal cDNA sequences mapped to the same gene (Table 1). Of the 1046 human genes examined on the microarray, the five most highly induced in heat-treated cells were heat shock protein 90 α (hsp90 α), dnaJ, hsp90 β , polyubiquitin, and t-complex polypeptide-1 (tcp-1) (Table 1). Three of the 17 clones did not match any entry in the public data base, though one of the clones (B7) exhibited significant homology to an EST from *Caenorhabditis elegans* (Table 1). Each of the novel sequences (B7-B9) exhibited ~ 2 -fold induction (Table 1) and relatively low-level expression (Table 2).

To confirm the microarray results, mRNA levels for each of the genes were measured by RNA blotting. Each of the genes that displayed heat shock induction, including the three novel

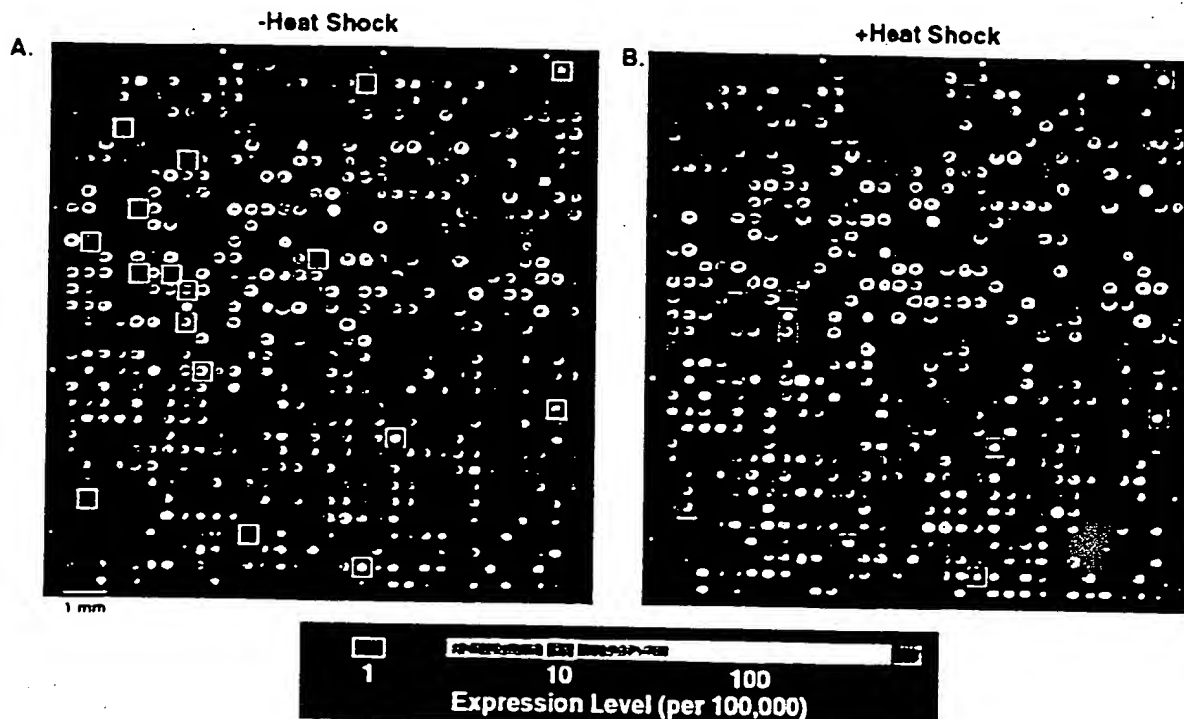


FIG. 1. Human gene expression monitored on a microarray. Fluorescent scans represented in a pseudocolor scale correspond to expression levels. The array contains 10 *Arabidopsis* controls (upper left corner, elements 1-10) and 1046 human peripheral blood cDNAs. Fluorescent probes were prepared by labeling mRNA from Jurkat cells grown at 37°C (-Heat Shock, A) or 43°C (+Heat Shock, B). Array elements that display altered fluorescence intensity (white boxes) corresponded to genes activated (red boxes) or repressed (green boxes) by heat shock. The color bar was calibrated in separate experiments using known quantities (wt/wt) of *Arabidopsis* control mRNAs added to the labeling reaction. Microarray rows (at left) and columns (at the top) are demarcated at 10 element increments (white circles). (Bar = 1 mm.)

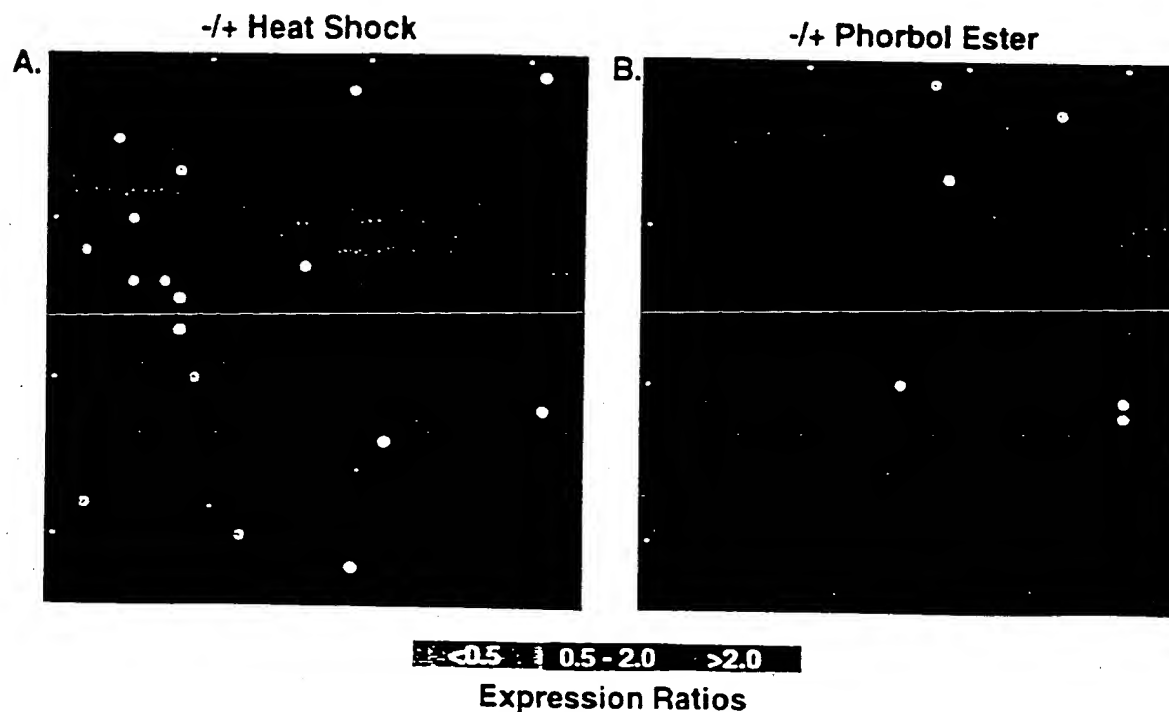


FIG. 2. Elemental displays of activated and repressed genes. Fluorescence ratios of two-color microarray scans (Fig. 1) are depicted schematically. Fluorescein-labeled probes from Jurkat cells subjected to (A) heat shock or (B) phorbol ester treatment were compared with Cy5-labeled probes from untreated cells. In a second set of reactions, the fluorescent groups were swapped (see text). The data represent the average of the ratios from two hybridizations, excluding values in which the difference of the two ratios was greater than half the average ratio. The color bar corresponds to expression ratios, which are independent of the absolute expression level of a given gene.

Table 1. Microarray elements corresponding to differentially expressed genes

Clone	Row	Column	Ratio	Blast identity	Accession no.
B1	24	21	0.5	CYC oxidase III	J01415, J01415
B2	1	31	0.5	β -Actin	NR, X00351
B3	15	8	0.5	CYC oxidase III	J01415, J01415
B4	32	19	0.5	CYC oxidase III	J01415, J01415
B5	17	8	0.5	CYC oxidase III	J01415, J01415
B6	22	31	0.5	β -Actin	NR, X00351
B7*	5	4	2.0	Novel†	U56653, U56654
B8	2	19	2.0	Novel†	U56655, U56656
B9	14	5	2.2	Novel†	U56657, U56658
B10	7	8	2.4	Polyubiquitin	X04803, X04803
B11	12	2	2.4	TCP-1	X52882, X52882
B12	28	2	2.5	Polyubiquitin	M17597, M17597
B13	14	7	2.5	Polyubiquitin	X04803, X04803
B14	20	9	2.6	HSP90 β	M16660, M16660
B15	30	12	4.0	DnaJ homolog	D13388, D13388
B16	10	5	5.8	HSP90 α	X07270, X07270
B17	13	16	6.3	HSP90 α	M27024, X15183
B18	7	19	2.0	β_2 -microglobulin	S54761, M30683
B19	21	30	2.1	Novel†	U56659, U56660
B20	3	26	2.2	β_2 -microglobulin	S54761, M30683
B21	1	18	2.6	PGK	M11968, L00160
B22	22	30	3.5	NF- κ B1	Z47744, M55643
B23	20	16	19	PAC-1	L11329, L11329

Clone name, array position (Fig. 1), fluorescence ratio, sequence identity, and accession number of cDNAs that manifested a differential expression pattern with probes prepared from heat shock- (B1-17) or phorbol ester-treated (B18-23) Jurkat cells. Clones showing >98% identity over 300 nucleotides were assumed to be identical to known sequences. All genes are nuclear except CYC oxidase III (mitochondrial). Accession numbers reflect the highest score for proximal and distal sequence traces, respectively. CYC, cytochrome c; TCP-1, T-complex polypeptide; HSP, heat shock protein; PGK, phosphoglycerate kinase; NF- κ B, nuclear factor-kappaB; PAC-1, phosphatase of activated cells; and NR, trace not readable due to the presence of poly(A)⁺ tract.

*B7 is 67% identical to an EST from *C. elegans* (D76026).

†No match in the public data bases.

Table 2. Human gene expression monitored by microarray and RNA blot analyses

Clone	Blast identity	Expression level, per 10 ⁴ mRNAs			
		Microarray	Ratio	RNA blot	Ratio
B1	CYC oxidase III	92/46	0.5	100/60	0.8
B2	β -Actin	240/120	0.5	270/280	1.0
B3	CYC oxidase III	36/18	0.5	ND	ND
B4	CYC oxidase III	76/38	0.5	ND	ND
B5	CYC oxidase III	62/31	0.5	ND	ND
B6	β -Actin	180/89	0.5	ND	ND
B7	Novel (weakly to D76026)	1.3/2.6	2.0	0.77/1.8	2.3
B8	Novel	2.0/4.0	2.0	1.5/3.4	2.3
B9	Novel	0.8/1.8	2.2	1.2/1.8	1.5
B10	Polyubiquitin	0.8/1.9	2.4	25/89	3.6
B11	TCP-1	2.3/5.5	2.4	7.1/27	3.8
B12	Polyubiquitin	0.8/2.0	2.5	ND	ND
B13	Polyubiquitin	1.7/4.3	2.5	ND	ND
B14	HSP90 β	75/200	2.6	30/120	4.0
B15	DnaJ homolog	1.0/4.0	4.0	1.6/13	8.1
B16	HSP90 α	0.6/3.5	5.8	3.2/29	9.1
B17	HSP90 α	0.8/5.0	6.3	8.6/62	7.2
B18	β_2 -microglobulin	1.0/2.0	2.0	5.4/15	2.8
B19	Novel	1.2/2.5	2.1	4.5/9.5	2.5
B20	β_2 -microglobulin	2.7/5.9	2.2	ND	ND
B21	Phosphoglycerate kinase	2.4/6.2	2.6	4.7/9.2	2.0
B22	NF-KB1	1.7/6.0	3.5	0.65/4.7	7.2
B23	PAC-1	0.5/9.5	19	0.21/15	71

Shown are expression levels per 100,000 mRNAs (wt/wt) of genes assayed with a microarray (Fig. 1) or RNA blot. Ratios correspond to values from cells subjected to heat shock (B1-17) or phorbol ester treatment (B18-23) relative to untreated cells. Clone and gene names are given in Table 1. ND, not determined.

sequences, exhibited elevated mRNA levels by dot blot analysis (Table 2). In all cases, expression ratios as determined by the two procedures differed by <2-fold for the genes identified in the heat shock experiments (Table 2). The two assays differed more widely in terms of assessing absolute expression levels; nonetheless, absolute expression as monitored on a microarray typically correlated with RNA blots to within a factor of five (Table 2).

Phorbol Ester Signaling. To explore a signaling pathway distinct from the heat shock response, microarrays were used to examine the cellular effects of phorbol ester treatment. Jurkat cells were treated with phorbol ester, harvested, lysed, and used as a source of mRNA. Samples of mRNA from untreated or phorbol ester-stimulated cells were labeled with reverse transcriptase. The probes were mixed, hybridized to microarrays, and scanned for fluorescence emission of the two fluorescent groups. A total of six array elements displayed ≥ 2.0 -fold elevated signals with probes from phorbol ester-treated cells relative to control samples (Fig. 2B).

To determine the identity of the phorbol ester-induced genes, clones corresponding to the six array elements were sequenced. Data base searches revealed perfect matches for five of the six sequences (Table 1). The two most highly induced genes were the PAC-1 tyrosine phosphatase and nuclear factor-kappa B1 (NF- κ B1); modest activation was observed for phosphoglycerate kinase and β_2 -microglobulin (Table 1). One remaining clone (B19) did not match any entry in the public data base (Table 1). B19 displayed a 2.1-fold induction and, similar to the novel heat shock genes, a relatively low absolute expression level (Tables 1 and 2). All six of the phorbol ester-inducible genes displayed increased steady-state mRNA levels by RNA blotting (Table 2). PAC-1 expression (Fig. 1; Table 2) defined a detection limit of $\sim 1:500,000$ for the assay.

Transcript Imaging in Human Tissues. To determine whether microarrays could be used to monitor expression in human tissues, probes were prepared from human bone mar-

row, brain, prostate, and heart by labeling each mRNA sample with Cy5-dCTP. In a separate reaction, a control probe was prepared by labeling Jurkat mRNA with fluorescein-dCTP. The four Cy5-labeled probes were each mixed with an aliquot of the fluorescein-labeled control sample, and the four mixtures were hybridized to separate microarrays. The arrays were washed and scanned for fluorescence emission, and hybridization signals for each of the tissues samples were normalized to the Jurkat control to generate an expression profile for each of the 1046 clones present on the array.

Detectable expression was observed for all 15 of the heat shock and phorbol ester-regulated genes in the four tissue types examined (Fig. 3). In general, the expression level of each gene in Jurkat cells correlated rather closely with expression in the four tissues (Table 2; Fig. 3). Genes encoding β -actin and cytochrome c oxidase, the two most highly expressed of the 15 genes in Jurkat cells (Table 2), were highly expressed in bone marrow, brain, prostate, and heart (Fig. 3A). Expression of cytochrome c oxidase, hsp90 α , and the novel B7 sequence was significantly greater in heart than in the other tissues (Fig. 3).

DISCUSSION

Many of the heat shock genes identified in this study encode factors that function either as molecular "chaperones" (HSP90 α , HSP90 β , DnaJ, TCP-1) or as mediators of protein degradation (polyubiquitin). The identification of these sequences is consistent with the biochemical basis of heat shock induction (10-15). Proteins undergo denaturation at elevated temperatures, and those that fail to maintain proper conformation must be selectively degraded (10-15). It will be interesting to determine whether the three novel heat shock-inducible sequences (B7-B9) mediate protein folding and turnover or possess some other biochemical activity. Complete nucleotide sequence determination, conceptual translation, expression monitoring, and biochemical analysis should provide a detailed functional understanding of these genes.

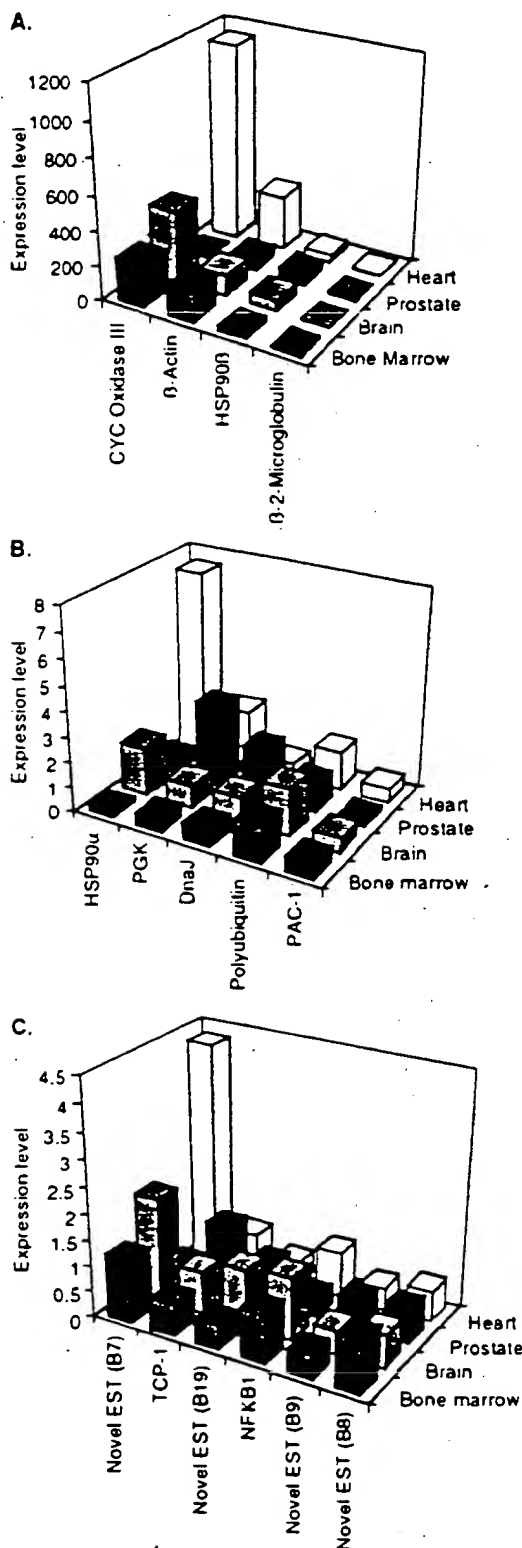


FIG. 3. Transcript profiles of heat shock and phorbol ester-regulated genes. Gene expression levels per 100,000 mRNAs (x-axes) are shown for 15 genes (Table 1) in human bone marrow (red), brain (green), prostate (blue), and heart (yellow). Genes are grouped according to expression levels (A–C).

Phorbol ester, a potent activator of protein kinase C (16, 17), induced a set of genes distinct from those involved in the heat shock pathway. The most highly induced gene identified in this study, *PAC-1*, encodes a nuclear tyrosine kinase that may play a role in regulating transcription and cell cycle progression (18). *NF-κB1*, a second phorbol ester-inducible gene, is an intensively studied member of the Rel transcription factor family (19–21). The Rel proteins are activated by a large number of stimuli, including phorbol esters, cytokines, bacterial and viral pathogens, and ultraviolet light (19–21). Modest activation was observed for three sequences not known to be inducible by phorbol esters, including phosphoglycerate kinase, β_2 -microglobulin, and a novel human gene (B19). Extensive expression monitoring with microarrays should assist in understanding how each of these genes integrate into the highly complex phorbol ester signaling pathway.

It is striking that four novel human genes were discovered with an array of 1000 randomly chosen clones, particularly because the heat shock and phorbol ester signaling pathways have been so intensively studied (10–21). The facile discovery of these sequences underscores the fact that microarrays can be used for gene discovery in the absence of any sequence information. By this approach, clones are chosen at random from any library of interest and only those clones that display interesting expression patterns are sequenced and characterized. This parallel assay, coupled with a modest DNA sequencing facility, allows high-throughput human genome expression analysis and gene discovery.

Genes that are activated or repressed by a given stimulus provide functional clues to the cellular pathway involved (22–24). Detailed examination of these gene expression "signatures" can provide a dynamic view of the mode of action of a given signaling substance (22–24). Microarrays may thus allow rapid mechanistic examination of hormones, drugs, elicitors, and other small molecules; moreover, functional analysis of transcription factors, kinases, growth factors, cytokines, receptors, and other gene products should be possible. Efforts are underway to develop mRNA amplification strategies to enable probe preparation from minute tissue samples. This capability might allow for high-throughput patient screening in a clinical setting.

The current detection limit of the assay allows monitoring of transcripts that represent $\sim 1:500,000$ (wt/wt) of the total mRNA. This 10-fold increase in sensitivity compared with the original report (4) was achieved largely by modifying the coupling chemistry, which reduced background fluorescence. The significance of this improvement is considerable in that approximately half the human genes identified in this study, including all four novel sequences, exhibited expression levels below the original detection limit of 1:50,000 (4).

The ability to detect 2-fold changes in expression was achieved by the use of two-color fluorescence in the labeling and detection schemes, digitized data collection, and custom software. The importance of this capability is underscored by the fact that nearly all of the genes examined here exhibited < 6 -fold changes in expression. The four novel genes, which showed ≤ 2.2 -fold activation, were probably overlooked in previous screens that used conventional differential expression techniques. It may be possible to further improve the precision of the microarray assay by the use of closely related fluorescent analogs, such as Cy3 and Cy5, in the labeling and hybridization reactions.

Microarrays offer a number of advantages over other potential high-capacity approaches to expression analysis. The chip-based approach enables small hybridization volumes, high array densities, and the use of fluorescence labeling and detection schemes. These features provide a set of performance specifications that are unattainable with filter-based approaches (25, 26). The use of cDNA clones provides hybridization specificity that is not readily attained with oligo-

nucleotide arrays (27-30). The parallel format of the assay provides a simultaneous differential expression readout for >1000 genes. This contrasts with sequencing-based methods, which require serial data collection for expression analysis (31, 32). A commercial source of cDNA microarrays would greatly speed the use of a chip-based approach to expression analysis.

The availability of large numbers of ESTs (3) provides a rich resource of human cDNA clones for microarraying. The >400,000 ESTs in the public data bases represent a significant subset of all human genes (3, 33). Microarrays of thousands of ESTs will provide a powerful analytical tool for future human gene expression studies. The ~100,000 genes in the human genome (2, 33) emphasize the need for microarrays of greater density. Attempts to improve microdeposition techniques are underway and should allow construction of arrays containing a complete set of human gene targets (<http://cmgm.stanford.edu/~schena/>). Microarrays of ~100,000 cDNA elements would allow expression monitoring of the entire human genome in a single hybridization. This capacity, coupled with detailed biochemical analysis of the individual gene products, would greatly speed the functional analysis of the human genome.

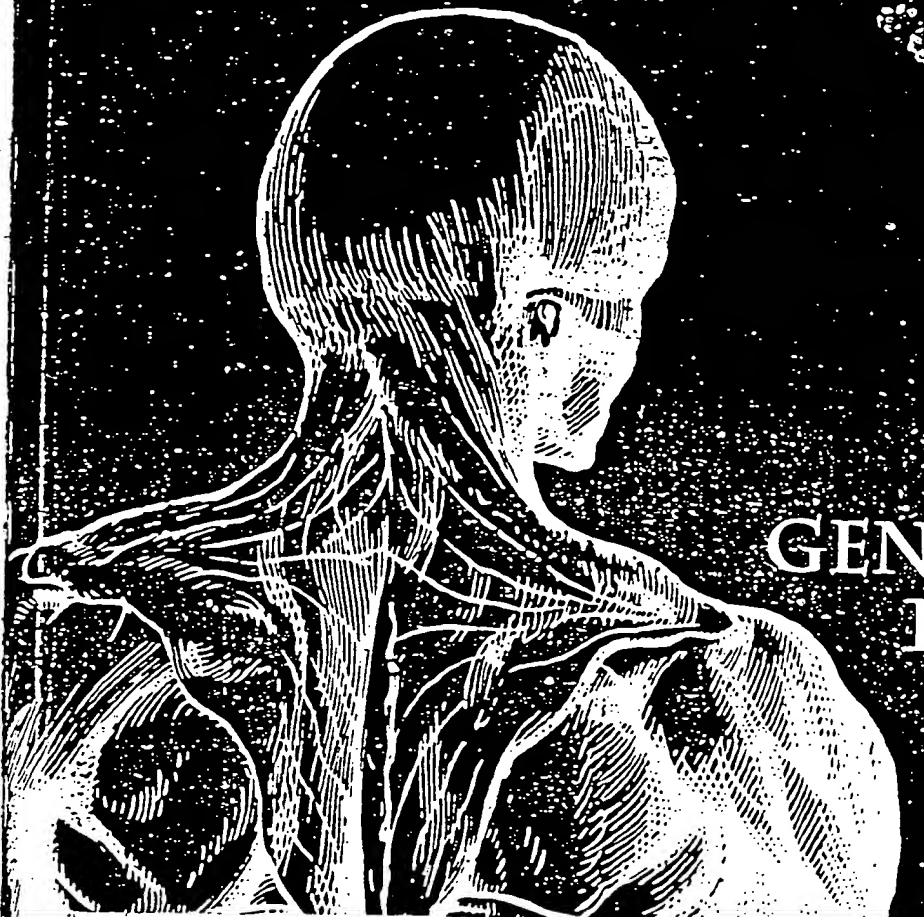
We thank S. Elledge (selledge@bcm.tmc.edu) for the human cDNA library, Qiagen representatives for help with plasmid purification, and A. J. Smith and colleagues at the Protein and Nucleic Acid (PAN) facility (Stanford) for oligonucleotide synthesis and DNA sequencing. We also thank members of the Davis, Brown, and Smith laboratories for critical comments and helpful discussions and Synteni employees for technical assistance. Support for R.W.D. was provided by the National Science Foundation (MCB9106011) and National Institutes of Health (R37HG00198) and for P.O.B. by the National Institutes of Health (3R21HG00450) and Howard Hughes Medical Institute. P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

1. Watson, J. D. (1993) *Gene* 135, 309-315.
2. Collins, F. S. (1995) *Proc. Natl. Acad. Sci. USA* 92, 10821-10823.
3. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., & Venter, J. C. (1991) *Science* 252, 1651-1656.
4. Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995) *Science* 270, 467-470.
5. Shalon, D. (1996) Ph.D. thesis (Stanford University).
6. Schena, M. (1996) *BioEssays* 18, 427-431.
7. Shalon, D., Smith, S. J., & Brown, P. O. (1996) *Genome Res.* 6, 639-645.
8. Schena, M., & Davis, R. W. (1994) *Proc. Natl. Acad. Sci. USA* 91, 8393-8397.
9. Schena, M., & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* 89, 3894-3898.
10. Jindal, S. (1996) *Trends Biotechnol.* 14, 17-20.
11. Wilkinson, K. D. (1995) *Annu. Rev. Nutr.* 15, 161-189.
12. Jakob, U., & Buchner, J. (1994) *Trends Biochem. Sci.* 19, 205-211.
13. Becker, J., & Craig, E. A. (1994) *Eur. J. Biochem.* 219, 11-23.
14. Cyr, D. M., Langer, T., & Douglas, M. G. (1994) *Trends Biochem. Sci.* 19, 176-181.
15. Craig, E. A., Weissman, J. S., & Horwich, A. L. (1994) *Cell* 78, 365-372.
16. Newton, A. C. (1995) *J. Biol. Chem.* 270, 28495-28498.
17. Nishizuka, Y. (1995) *FASEB J.* 9, 484-496.
18. Rohan, P. J., Davis, P., Moskaluk, C. A., Kearns, M., Krutzsch, H., Siebenlist, U., & Kelly, K. (1993) *Science* 259, 1762-1766.
19. Thanos, D., & Maniatis, T. (1995) *Cell* 80, 529-532.
20. Baeuerle, P. A., & Henkel, T. (1994) *Annu. Rev. Immunol.* 12, 141-179.
21. Liou, H.-C., & Baltimore, D. (1993) *Curr. Opin. Cell Biol.* 5, 477-487.
22. Cohen, G. B., Ren, R., & Baltimore, D. (1995) *Cell* 80, 237-248.
23. Chan, A. C., Desai, D. M., & Weiss, A. (1994) *Annu. Rev. Immunol.* 12, 555-592.
24. Crabtree, G. R., & Clipstone, N. A. (1994) *Annu. Rev. Biochem.* 63, 1045-1083.
25. Gress, T. M., Hoheisel, J. D., Lennon, G. G., Zehetner, G., & Lehrach, H. (1992) *Mamm. Genome* 3, 609-619.
26. Bernard, K., Auphan, N., Granjeaud, S., Victorero, G., Schmitt-Verhulst, A.-M., Jordan, B. R., & Nguyen, C. (1996) *Nucleic Acids Res.* 24, 1435-1442.
27. Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., & Solas, D. (1991) *Science* 251, 767-773.
28. Southern, E. M., Maskos, U., & Elder, J. K. (1992) *Genomics* 13, 1008-1017.
29. Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R., & Smith, L. M. (1994) *Nucleic Acids Res.* 22, 5456-5465.
30. Matson, R. S., Rampal, J., Pentoney, S. L., Jr., Anderson, P. D., & Coassin, P. (1995) *Anal. Biochem.* 224, 110-116.
31. Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995) *Science* 270, 484-487.
32. Adams, M. D. (1996) *BioEssays* 18, 261-262.
33. Fields, C., Adams, M. D., White, O., & Venter, J. C. (1994) *Nat. Genet.* 7, 345-346.

S
SCI08
Dup

SCIENTIFIC AMERICAN

20 OCTOBER 1995
Vol. 273 • Pages 349-504



GENOME
ISSUE

*****J-DIGIT 510
80258222281012#AS 12/01/95 E 5235 74
ELDO C KOENIG
35005 M FAIRVIEW RD
OCNOMOC
MI 53066-3312

COVER

The Genome Project adds a new dimension to questions on gene expression in humans and model systems. A chart on page 415 summarizes progress in the *Caenorhabditis elegans* Genome Project and indicates some ways information about sequences can be used.

News stories, Articles, Perspectives, Policy Forums, and Reports focus on technological developments, clinical applications, and ethical concerns resulting from the burgeoning of genomic information. [*C. elegans* image: F. Maduro and D. Pilgrim, University of Alberta]



REPORTS

Cosmogenic Ages for Earthquake Recurrence Intervals and Debris Flow Fan Deposition, Owens Valley, California 447
P. R. Bierman, A. R. Gillespie, M. W. Caffee

Lithoautotrophic Microbial Ecosystems in Deep Basalt Aquifers 450
T. O. Stevens and J. P. McKinley

Large Arctic Temperature Change at the Wisconsin-Holocene Glacial Transition 455
K. M. Cuffey, G. D. Clow, R. B. Alley, M. Stuiver, E. D. Waddington, R. W. Saltus

Superplasticity in Earth's Lower Mantle: Evidence from Seismic Anisotropy and Rock Physics 458
S.-i. Karato, S. Zhang, H.-R. Wenk

Large-Scale Interplanetary Magnetic Field Configuration Revealed by Solar Radio Bursts 461
M. J. Reiner, J. Fainberg, R. G. Stone

Role of Yeast Insulin-Degrading Enzyme Homologs in Propheromone Processing and Bud Site Selection 464
N. Adames, K. Blundell, M. N. Ashby, C. Boone

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray 467
M. Schena, D. Shalon, R. W. Davis, P. O. Brown

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients 470
C. Bordignon, L. D. Notarangelo, N. Nobili, G. Ferrari, G. Casorati, P. Panina, E. Mazzolari, D. Maggioni, C. Rossi, P. Servida, A. G. Ugazio, F. Mavilio

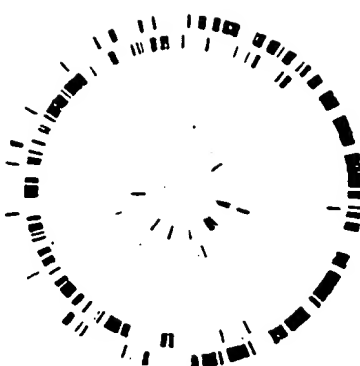
T Lymphocyte-Directed Gene Therapy for ADA⁻ SCID: Initial Trial Results After 4 Years 475
R. M. Blaese, K. W. Culver, A. D. Miller, C. S. Carter, T. Fleisher, M. Clerici, G. Shearer, L. Chang, Y. Chiang, P. Tolstoshev, J. J. Greenblatt, S. A. Rosenberg, H. Klein, M. Berger, C. A. Mullen, W. J. Ramsey, L. Muul, R. A. Morgan, W. F. Anderson

Physical Map and Organization of *Arabidopsis thaliana* Chromosome 4 480
R. Schmidt, J. West, K. Love, Z. Lenehan, C. Lister, H. Thompson, D. Bouchez, C. Dean

Serial Analysis of Gene Expression 484
V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler

TECHNICAL COMMENTS

The Radius of Gyration of an Apomyoglobin Folding Intermediate 487
D. Eliezer, P. A. Jennings, P. E. Wright, S. Doniach, K. O. Hodgson, H. Tsuruta



397
Good things in small genomes

AAS Board of Directors ☒ Indicates accompanying feature

Isaac J. Ayala
Retiring President
Chairman
Is R. Colwell
President
De Lubchenco
President-elect

Anne C. Roosevelt
Alan Schnesheim
Jean E. Taylor
Chang-Lin Tian
Nancy S. Wessler

William T. Golden
Treasurer
Richard S. Nicholson

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1233 H Street, NW, Washington, DC 20005. Second-class postage (publication No. 484450) paid at Washington, DC, and additional mailing offices. Copyright © 1995 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership and subscription (\$1 issues): \$97 (\$50 allocated to subscription). Domestic institutional subscription (\$1 issues): \$228. Foreign postage extra: Mexico, Caribbean (surface mail) \$53; other countries (air assist delivery) \$93. First class, airmail, student and emeritus rates on request. Canadian rates with GST available on request. GST #R123047800.

Change of address: allow 4 weeks, giving old and new addresses and 8-digit account number. Postmaster: Send change of address to Science, P.O. Box 18111, Danbury, CT 06813-1811. Single copy sales: \$7.00 per issue prepaid includes surface postage; bulk rates on request. Authorization to photocopy material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that \$3.00 per article is paid directly to CCC, 27 Congress Street, Salem, MA 01970. The identification code for Science is 0036-8075/95 \$3.00. Science is indexed in the Reader's Guide to Periodical Literature.

Ad1p sequence following Ser²⁰⁸ and occurs within the domain of Ad1p that shows homology with NDE (14). To delete the complete STE23 sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCAT-TCTTGCTCATTTTGATATTGCTC- TGATAGATTG-TACTGAGAGTGAC-3'; and 5'-GCTACAAACAGC-GTCGACTTGAATGCCCGGACATCTTGCAGTGT-GCGGTATTTACACCG-3') were used to amplify the *URA3* sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb Sal I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb Hpa I-Xho I fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb Pml I-Ecl136 II fragment of STE23, which occurs within a 6.2-kb Hind III-Bgl II genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.6-kb Bam HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Sterne, J. Thomer, *EMBO J.* 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI-Sst I fragment from pAXL1. Substitution mutations of the proposed active site of Ad1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (*axl1-H68A*, 5'-GTGCTCACAAGGCGT-GCCAAACCGGC-3'; *axl1-E71A*, 5'-AAGAATCAT-GTGGCACAAGGTGGCGC-3'; and *axl1-E71D*, 5'-AAGAATCATGTGATCACAAGGTGGCGC-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI-Msc I fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (*axl1-H68A*), p130 (*axl1-E71A*), and p132 (*axl1-E71D*). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI-Msc I fragment, to generate p161 (*axl1-E71A*), p162 (*axl1-*

H68A), and p163 (*axl1-E71D*).
32. We thank J. Becker and S. Michaels for providing a-factor antibodies; S. Michaels for discussing unpublished results and helping with the pulse-chase experiments; J. Brown, J. Chant, and S. Sanoers for their input concerning bud site selection experiments; M. Raymond, F. Tamanoi, and M. Whiteway for plasmids; M. Marra for providing the STE23 genomic fragment; and H. Bussey, J. Brown, N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 12 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (4) by a single round of reverse transcription (5). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thomer, *J. Cell Biol.* 120, 1203 (1993); R. Koling and C. P. Hollenberg, *EMBO J.* 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaels, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A.* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); _____ and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (*MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 sst1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1::URA3 ste23Δ::URA3*), Y221 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). *MATa* derivatives of SY2625 included the following strains: Y199 (SY2625 made *MATa*), Y278 (*ste22-1*), Y195 (*mfa1Δ::LEU2*), Y196 (*axl1Δ::LEU2*), and Y197 (*axl1::URA3*). The EG123 (*MATa leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). *MATa* derivatives of EG123 included the following strains: Y214 (EG123 made *MATa*) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate *MATa ste23* and *MATa axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the Bgl II

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of $\sim 1:50,000$. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig. 1, C

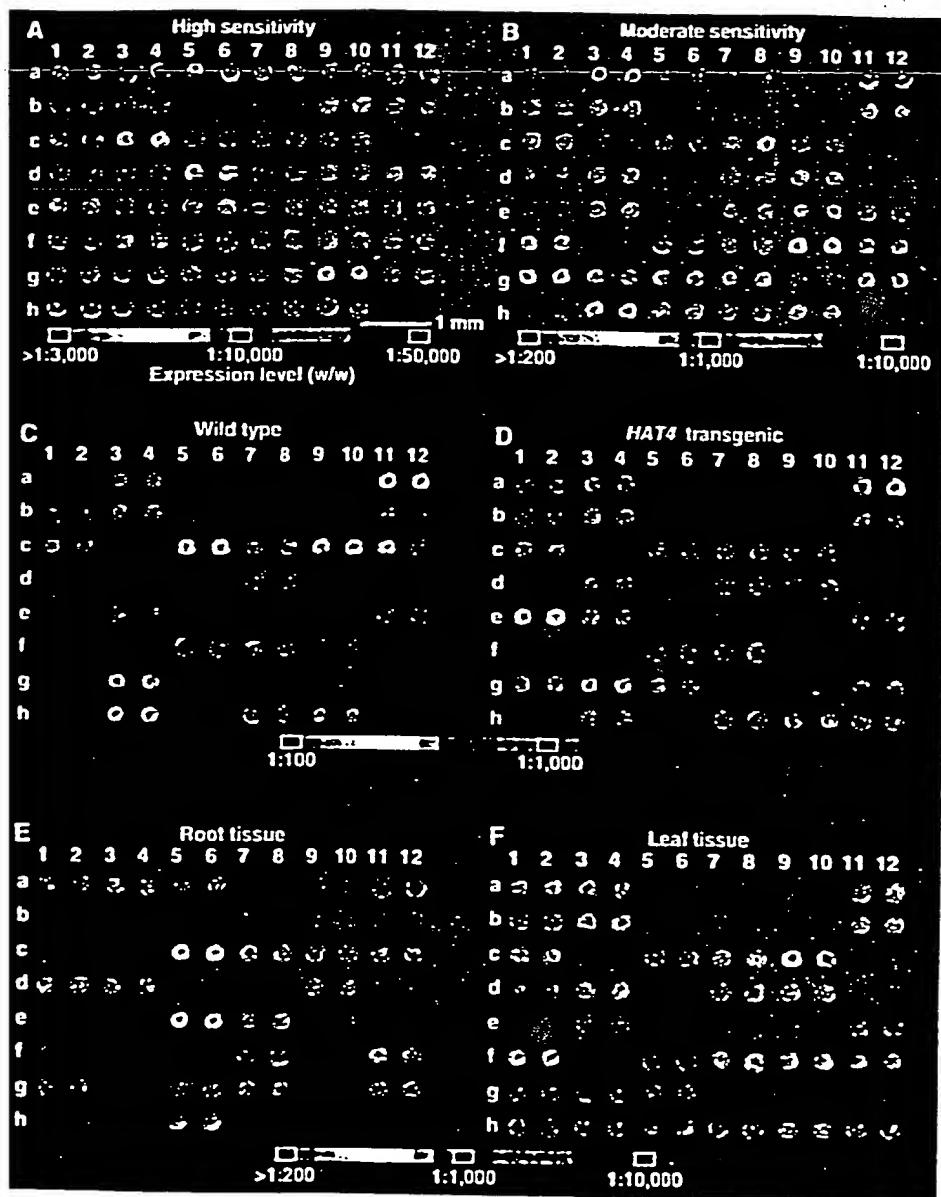


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained in the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization of fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the

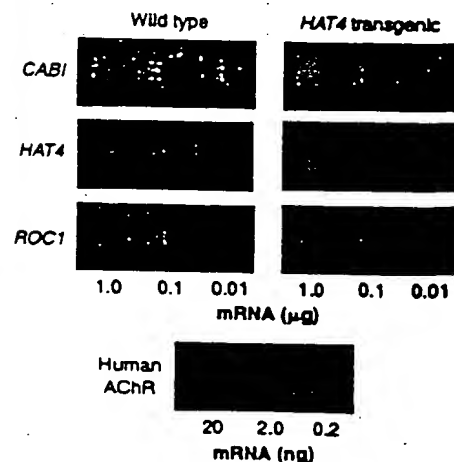


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated *CAB1* gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The *HAT4*-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for *HAT4*, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon *HAT4* overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and *HAT4*-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (2). The availability of 20,274 ESTs from *Arabidopsis* (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of *Arabidopsis* genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive *in vivo* sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	.
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	<i>CAB1</i>	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	GA4	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	<i>GBF-1</i>	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	<i>GBF-2</i>	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	<i>HAT1</i>	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	<i>HAT2</i>	Homeobox-leucine zipper 2	U09335
e1, 2	<i>HAT4</i>	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	<i>HAT5</i>	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	<i>HAT22</i>	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	<i>KNAT1</i>	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	<i>PPH1</i>	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	<i>ROC1</i>	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	<i>SAR1</i>	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	<i>TRP4</i>	Yeast tryptophan biosynthesis	X04273

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, *HAT4*-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
<i>CAB1</i>	1:48	1:83
<i>CAB1</i> (tg)	1:120	1:150
<i>HAT4</i>	1:8300	1:8300
<i>HAT4</i> (tg)	1:150	1:210
<i>ROC1</i>	1:1200	1:1800

REFERENCES AND NOTES

- urrent EST database (dbEST release 091495) the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries including 255,645 from the human genome and 1,044 from Arabidopsis. Access is available via World Wide Web (<http://www.ncbi.nlm.nih.gov>).
- Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1986); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 169 (1986); I. Hwang et al., *Plant J.* 1, 367 (1991); vis et al., *Plant Mol. Biol.* 24, 585 (1994); L. Le et al., *Mol. Gen. Genet.* 245, 390 (1994).
- ation, thesis, Stanford University (1995); ———, O. Brown, in preparation. Microarrays were printed on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine with one printing tip. The tip loaded 1 µl of PCR product (0.5 mg/ml) from 96-well microtiter plates deposited ~0.005 µl per slide on 40 slides at a spacing of 500 µm. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 4°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer containing 50% 1-methyl-2-pyrrolidone and 50% acetic acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser scanner that contained a computer-controlled XY stage and a microscope objective. A mixed argon-ion laser allowed sequential excitation of different fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 16-bit analog-to-digital board. Additional details of array fabrication and use may be obtained by e-mail (pbrown@cmgm.stanford.edu).
- Musubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
- enyated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with StrataScript RT-PCR kit (Stratagene) modified as follows: 50-µl reactions contained 0.1 µg/µl of total RNA, 0.1 ng/µl of human AChR cDNA, 0.05 µg/µl of oligo(dT) (21-mer), 1× first buffer, 0.03 U/µl of ribonuclease block, 500 U/µl oxadenosine triphosphate (dATP), 500 µM dUTP, 400 U/µl oxycytosine triphosphate (dCTP), 40 µM fluorimide-12-dCTP (or isosamine-5-dCTP), and 0.03 U/µl StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with 100% ethanol, and resuspended in 10 µl of TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 µl of 10 N NaOH and incubating for 10 min at 37°C. The sample was neutralized by addition of 2.5 µl of 1 M HCl (pH 8.0) and 0.25 µl of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 µl of H₂O, and reduced to 3.0 µl in a vacuum. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
- ization reactions contained 1.0 µl of fluorescently labeled nucleotide (5) and 1.0 µl of hybridization buffer (10× saline sodium citrate (SSC) and 0.2% SDS). The 2.0-µl probe mixtures were aliquoted onto the array surface and covered with cover slips (round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 42°C. Arrays were washed for 5 min at room temperature in low-stringency wash buffer (1× SSC, 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1× SSC and 0.1% SDS) and scanned in 0.1× SSC with the use of a resonance laser-scanning device (3).
- of poly(A)⁺ mRNA (4, 5) were spotted onto membranes (Nyttran) and crosslinked with ultraviolet light with the use of a Stratamaker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridization was carried out
- facturer. Quantitation was performed on a PhosphorImager (Molecular Dynamics).
- B. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
- H. Hofte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).
- N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Betancourt-Charlton, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).
- E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
- The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Muligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00196 (R.W.D.) and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)-identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector-mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells in an engraftment model.

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.